

Lecture Notes in Artificial Intelligence 6291

Edited by R. Goebel, J. Siekmann, and W. Wahlster

Subseries of Lecture Notes in Computer Science

Yaxin Bi Mary-Anne Williams (Eds.)

# Knowledge Science, Engineering and Management

4th International Conference, KSEM 2010  
Belfast, Northern Ireland, UK, September 1-3, 2010  
Proceedings



Springer

Series Editors

Randy Goebel, University of Alberta, Edmonton, Canada  
Jörg Siekmann, University of Saarland, Saarbrücken, Germany  
Wolfgang Wahlster, DFKI and University of Saarland, Saarbrücken, Germany

Volume Editors

Yaxin Bi  
University of Ulster, School of Computing and Mathematics  
Newtownabbey, Co. Antrim, BT37 0QB, UK  
E-mail: y.bi@ulster.ac.uk

Mary-Anne Williams  
University of Technology, Faculty of Information Technology  
Sydney, NSW 2007, Australia  
E-mail: Mary-Anne@it.uts.edu.au

Library of Congress Control Number: 2010932504

CR Subject Classification (1998): I.2.4, H.3, I.2, H.4, J.1, H.2.8

LNCS Sublibrary: SL 7 – Artificial Intelligence

ISSN 0302-9743

ISBN-10 3-642-15279-1 Springer Berlin Heidelberg New York  
ISBN-13 978-3-642-15279-5 Springer Berlin Heidelberg New York

This work is subject to copyright. All rights are reserved, whether the whole or part of the material is concerned, specifically the rights of translation, reprinting, re-use of illustrations, recitation, broadcasting, reproduction on microfilms or in any other way, and storage in data banks. Duplication of this publication or parts thereof is permitted only under the provisions of the German Copyright Law of September 9, 1965, in its current version, and permission for use must always be obtained from Springer. Violations are liable to prosecution under the German Copyright Law.

springer.com

© Springer-Verlag Berlin Heidelberg 2010  
Printed in Germany

Typesetting: Camera-ready by author, data conversion by Scientific Publishing Services, Chennai, India  
Printed on acid-free paper 06/3180

# Preface

The proceedings contains the papers selected for the 4th International Conference on Knowledge Science, Engineering and Management (KSEM). The collection of papers represent the wide range of research activities that have been carried out, covering knowledge representation and reasoning, knowledge extraction, knowledge integration, data mining and knowledge discovery, etc. Additionally, two special sessions: “Theory and Practice of Ontology for the Semantic Web” and “Application of Data Mining to Seismic Data Analysis for Earthquake Study” have been included along with the main conference program. All of the papers have been peer reviewed by the members of the Program Committee and external reviewers; 45 regular papers and 12 short papers were selected. They represent the state of the art of research in KSEM-related research areas.

The International Conference on Knowledge Science, Engineering and Management in 2010 was the fourth event in the series of successful conferences, previously held in Guilin, China, Melbourne, Australia, and Vienna, Austria. It specifically focused on becoming a premier forum for prototype and deployed knowledge engineering and knowledge-based systems. KSEM offers an exceptional opportunity for presenting original work, the latest scientific and technological advances on knowledge-related systems, and discussing and debating practical challenges and opportunities for the research community.

A great many people contributed to making KSEM2010 a successful conference. First of all, we would like to thank the KSEM2010 Organizing Committee and the School of Computing and Mathematics and the University of Ulster for providing crucial support throughout the organization of the conference. Secondly, we are immensely grateful for the financial support donated by the sponsors of the conference, including the European Office of Aerospace Research and Development, United States Air Force Research Laboratory, *Artificial Intelligence Journal*, Belfast Visitor & Convention Bureau, Springer, and the Computer Science Research Institute, University of Ulster. Additionally, we would like to thank the General Conference Chairs, Bryan Scotney from the University of Ulster, Jin Zhi from Peking University of China, the Steering Committee Chairs, Chengqi Zhang from the University of Technology, Sydney and Ruqian Lu from the Chinese Academy of Science, for their support and assistance. Last, but not least, we would like to express our gratitude to the authors who submitted papers to the conference, the Programme Committee, and the additional reviewers who played a significant role in the review process.

June 2010

Yaxin Bi  
Mary-Anne Williams

# Conference Organization

## Programme Chairs

Yaxin Bi  
Mary-Anne Williams

## Programme Committee

Harith Alani	Wei Liu
Andreas A Albrecht	Weiru Liu
Klaus Dieter Althoff	James Lu
Salem Benferhat	Ronald Maier
Philippe Besnard	Pierre Marquis
Gernhard Brewka	Stewart Massie
Krysia Broda	Paul McKeivitt
Cungen Cao	John-Jules Meyer
Liming Chen	Michele Missikoff
Paolo Ciancarini	Riichiro Mizoguchi
Ruth Cobos Pérez	Leora Morgenstern
Ireneusz Czarnowski	Bernhard Nebel
Richard Dapoigny	Oleg Oku
Stefan Decker	Dan O'Leary
Xiaotie Deng	Barry O'Sullivan
Juan Manuel Dodero	Maurice Pagnucco
Josep Domènech	Pavlos Peppas
Josep Domingo-Ferrer	Ulrich Reimer
Wilfried Grossmann	Werner Retschitzegger
Gongde Guo	Bodo Rieger
Udo Hahn	Juan Jose Rodriguez
Achim Hoffmann	Maria Dolores Rodriguez-Moreno
Joana Hois	Michael Thielscher
Jun Hong	A Min Tjoa
Eyke Hüllermeier	Gianluca Torta
Anthony Hunter	Eric Tsui
Van-Nam Huynh	Carl Vogel
Takayuki Ito	Kewen Wang
Byeong Ho Kang	Rosina Weber
Gabriele Kern-Isberner	Eoin Whelan
John Kidd	Glenn Wightwick
Konstantinos Kotis	Qingxiang Wu
Jerome Lang	Wang Xun

Takahira Yamaguchi  
Jia-Huai You  
Slawomir Zadrozny  
Zdenek Zdrahal  
Qingtian Zeng  
Chunxia Zhang

Shichao Zhang  
Songmao Zhang  
Zhi-Hua Zhou  
Meiyun Zuo

## External Reviewers

Kerstin Bach  
Georgeta Bordea  
Arnau Erola Cañellas  
Katsuhide Fujita  
David Glass  
Sid Gunawardena  
Wei Hu  
Anna jurek  
Xiaoshan Li  
Xiaoshan Li  
Yibing Luo  
Takeshi Morita  
Kedian Mu  
Amedeo Napoli  
Amedeo Napoli

Regis Newo  
Antonio De Nicola  
Guilin Qi  
Jian Feng Shi  
Fabrizio Smith  
Pengfei Sun  
Amirreza Tahamtan  
Ilya Waldstein  
Haiying Wang  
Puwei Wang  
Stefan Woelfl  
Shengli Wu  
Qian Yang

## Sponsoring Institutions

European Office of Aerospace Research and Development, United States Air  
Force Research Laboratory  
Artificial Intelligence Journal  
Belfast Visitor & Convention Bureau, Northern Ireland, UK  
Springer  
Computer Science Research Institute, Faculty of Computing and Engineering,  
University of Ulster, Northern Ireland, UK

# Table of Contents

Mining Video Data: Learning about Activities . . . . .	1
<i>Anthony G. Cohn</i>	
Ontology Languages and Engineering . . . . .	2
<i>Ian Horrocks</i>	
Theory of Belief Functions for Data Analysis and Machine Learning Applications: Review and Prospects . . . . .	3
<i>Thierry Denoeux</i>	
Modeling Ontological Concepts of Locations with a Heterogeneous Cardinal Direction Model . . . . .	4
<i>Hui Shi and Yohei Kurata</i>	
Finding Minimal Rare Itemsets and Rare Association Rules . . . . .	16
<i>Laszlo Szathmary, Petko Valtchev, and Amedeo Napoli</i>	
Background Knowledge Integration in Clustering Using Purity Indexes . . . . .	28
<i>Germain Forestier, Cédric Wemmert, and Pierre Gançarski</i>	
A Comparison of Merging Operators in Possibilistic Logic . . . . .	39
<i>Guilin Qi, Weiru Liu, and David Bell</i>	
Modelling and Reasoning in Metamodelling Enabled Ontologies . . . . .	51
<i>Nophadol Jekjantuk, Gerd Gröner, and Jeff. Z. Pan</i>	
Towards Encoding Background Knowledge with Temporal Extent into Neural Networks . . . . .	63
<i>Han The Anh and Nuno C. Marques</i>	
A Fuzzy Description Logic with Automatic Object Membership Measurement . . . . .	76
<i>Yi Cai and Ho-Fung Leung</i>	
Decomposition-Based Optimization for Debugging of Inconsistent OWL DL Ontologies . . . . .	88
<i>Jianfeng Du and Guilin Qi</i>	
A Concept Hierarchy Based Ontology Mapping Approach . . . . .	101
<i>Ying Wang, Weiru Liu, and David Bell</i>	
Composing Cardinal Direction Relations Basing on Interval Algebra . . . .	114
<i>Juan Chen, Haiyang Jia, Dayou Liu, and Changhai Zhang</i>	

Retrieval Result Presentation and Evaluation . . . . .	125
<i>Shengli Wu, Yaxin Bi, and Xiaoqin Zeng</i>	
Autonomy: Life and Being . . . . .	137
<i>Mary-Anne Williams</i>	
A Method of Social Collaboration and Knowledge Sharing Acceleration for e-Learning System: The Distance Learning Network Scenario . . . . .	148
<i>Przemysław Różewski</i>	
A Comparative Study of Target-Based Evaluation of Traditional Craft Patterns Using Kansei Data . . . . .	160
<i>Van-Nam Huynh, Yoshiteru Nakamori, and Hongbin Yan</i>	
Optimization of Multiple Related Negotiation through Multi-Negotiation Network . . . . .	174
<i>Fenghui Ren, Minjie Zhang, Chunyan Miao, and Zhiqi Shen</i>	
Reasoning Activity for Smart Homes Using a Lattice-Based Evidential Structure . . . . .	186
<i>Jing Liao, Yaxin Bi, and Chris Nugent</i>	
Knowledge Modelling to Support Inquiry Learning Tasks . . . . .	198
<i>Annika Wolff, Paul Mulholland, Zdenek Zdrahal, and Miroslav Blasko</i>	
Building the Knowledge Base to Support the Automatic Animation Generation of Chinese Traditional Architecture . . . . .	210
<i>Gongjin Wei, Weijing Bai, Meifang Yin, and Songmao Zhang</i>	
Discovery of Relation Axioms from the Web . . . . .	222
<i>Luis Del Vasto Terrientes, Antonio Moreno, and David Sánchez</i>	
An Improved Piecewise Aggregate Approximation Based on Statistical Features for Time Series Mining . . . . .	234
<i>Chonghui Guo, Hailin Li, and Donghua Pan</i>	
Incorporating Duration Information in Activity Recognition . . . . .	245
<i>Priyanka Chaurasia, Bryan Scotney, Sally McClean, Shuai Zhang, and Chris Nugent</i>	
A Graphical Model for Risk Analysis and Management . . . . .	256
<i>Xun Wang and Mary-Anne Williams</i>	
Towards Awareness Services Usage Characterization: Clustering Sessions in a Knowledge Building Environment . . . . .	270
<i>Pedro G. Campos and Ruth Cobos</i>	



Adjusting Class Association Rules from Global and Local Perspectives Based on Evolutionary Computation . . . . .	282
<i>Guangfei Yang, Jiangning Wu, Shingo Mabu, Kaoru Shimada, and Kotaro Hirasawa</i>	
Probabilistic Declarative Process Mining . . . . .	292
<i>Elena Bellodi, Fabrizio Riguzzi, and Evelina Lamma</i>	
Making Ontology-Based Knowledge and Decision Trees Interact: An Approach to Enrich Knowledge and Increase Expert Confidence in Data-Driven Models . . . . .	304
<i>Iyan Johnson, Joël Abécassis, Brigitte Charnomordic, Sébastien Destercke, and Rallou Thomopoulos</i>	
A Novel Initialization Method for Semi-supervised Clustering . . . . .	317
<i>Yanzhong Dang, Zhaoguo Xuan, Lili Rong, and Ming Liu</i>	
Constructing and Mapping Fuzzy Thematic Clusters to Higher Ranks in a Taxonomy . . . . .	329
<i>Boris Mirkin, Susana Nascimento, Trevor Fenner, and Luís Moniz Pereira</i>	
Anticipation as a Strategy: A Design Paradigm for Robotics . . . . .	341
<i>Mary-Anne Williams, Peter Gärdenfors, Benjamin Johnston, and Glenn Wightwick</i>	
Modular Logic Programming for Web Data, Inheritance and Agents . . . .	354
<i>Isambo Karali</i>	
Automatic Collecting Technique of Low Frequency Electromagnetic Signals and Its Application in Earthquake Study . . . . .	366
<i>Xuemin Zhang, Roberto Battiston, Xuhui Shen, Zhima Zeren, Xinyan Ouyang, Jiadong Qian, Jing Liu, Jianping Huang, and Yuanqing Miao</i>	
Improving Search in Tag-Based Systems with Automatically Extracted Keywords . . . . .	378
<i>Ruba Awawdeh and Terry Anderson</i>	
Towards a Framework for Trusting the Automated Learning of Social Ontologies . . . . .	388
<i>Konstantinos Kotis, Panos Alexopoulos, and Andreas Papasalouros</i>	
PlayPhysics: An Emotional Games Learning Environment for Teaching Physics . . . . .	400
<i>Karla Muñoz, Paul Mc Kevitt, Tom Lunney, Julieta Noguez, and Luis Neri</i>	

A SOM-Based Technique for a User-Centric Content Extraction and Classification of Web 2.0 with a Special Consideration of Security Aspects . . . . .	412
<i>Amirreza Tahamtan, Amin Anjomshoaa, Edgar Weippl, and A. Min Tjoa</i>	
Modularizing Spatial Ontologies for Assisted Living Systems . . . . .	424
<i>Joana Hois</i>	
Towards Scalable Instance Retrieval over Ontologies . . . . .	436
<i>Alissa Kaprunova, Ralf Möller, Sebastian Wandelt, and Michael Wessel</i>	
MindDigger: Feature Identification and Opinion Association for Chinese Movie Reviews . . . . .	449
<i>Lili Zhao and Chunping Li</i>	
The Impact of Latency on Online Classification Learning with Concept Drift . . . . .	459
<i>Gary R. Marrs, Ray J. Hickey, and Michaela M. Black</i>	
Efficient Reasoning with RCC-3D . . . . .	470
<i>Julia Albath, Jennifer L. Leopold, Chaman L. Sabharwal, and Kenneth Perry</i>	
Automated Ontology Generation Using Spatial Reasoning . . . . .	482
<i>Alton Coalter and Jennifer L. Leopold</i>	
Facilitating Experience Reuse: Towards a Task-Based Approach . . . . .	494
<i>Ying Du, Liming Chen, Bo Hu, David Patterson, and Hui Wang</i>	
Behavioural Rule Discovery from Swarm Systems . . . . .	506
<i>David Stoops, Hui Wang, George Moore, and Yaxin Bi</i>	
Knowledge Discovery Using Bayesian Network Framework for Intelligent Telecommunication Network Management . . . . .	518
<i>Abul Bashar, Gerard Parr, Sally McClean, Bryan Scotney, and Detlef Nauck</i>	
Combining Logic and Probabilities for Discovering Mappings between Taxonomies . . . . .	530
<i>Rémi Tournaire, Jean-Marc Petit, Marie-Christine Rousset, and Alexandre Termier</i>	
An Ontology-Based Semantic Web Service Space Organization and Management Model . . . . .	543
<i>Kun Yang and Zhongzhi Shi</i>	

Genetic Algorithm-Based Multi-objective Optimisation for QoS-Aware Web Services Composition .....	549
<i>Li Li, Pengyi Yang, Ling Ou, Zili Zhang, and Peng Cheng</i>	
Knowledge Merging under Multiple Attributes .....	555
<i>Bo Wei, Zhi Jin, and Didar Zowghi</i>	
Feature Selection Based on Mutual Information and Its Application in Hyperspectral Image Classification .....	561
<i>Na Yao, Zongjian Lin, and Jingxiong Zhang</i>	
Static, Dynamic and Semantic Dimensions: Towards a Multidisciplinary Approach of Social Networks Analysis .....	567
<i>Christophe Thovex and Francky Trichet</i>	
Knowledge Based Systems and Metacognition in Radar .....	573
<i>Gerard T. Capraro and Michael C. Wicks</i>	
Maximus-AI: Using Elman Neural Networks for Implementing a SLMR Trading Strategy .....	579
<i>Nuno C. Marques and Carlos Gomes</i>	
A Formalism for Causal Explanations with an Answer Set Programming Translation .....	585
<i>Yves Moinard</i>	
Earthquake Prediction Based on Levenberg-Marquardt Algorithm Constrained Back-Propagation Neural Network Using DEMETER Data .....	591
<i>Lingling Ma, Fangzhou Xu, Xinhong Wang, and Lingli Tang</i>	
Affinity Propagation on Identifying Communities in Social and Biological Networks .....	597
<i>Caiyan Jia, Yawen Jiang, and Jian Yu</i>	
Semantic Decomposition of Indicators and Corresponding Measurement Units .....	603
<i>Michaela Denk and Wilfried Grossmann</i>	
Engineering Knowledge for Assistive Living .....	609
<i>Liming Chen and Chris Nugent</i>	
Large-Scale, Exhaustive Lattice-Based Structural Auditing of SNOMED CT .....	615
<i>Guo-Qiang Zhang</i>	
<b>Author Index</b> .....	617

# Mining Video Data: Learning about Activities

Anthony G. Cohn

School of Computing  
University of Leeds, UK

In this talk I will present ongoing work at Leeds on building models of video activity. I will present techniques, both supervised and unsupervised, for learning the spatio-temporal structure of tasks and events from video or other sensor data. In both cases, the representation will exploit qualitative spatio-temporal relations. A novel method for robustly transforming video data to qualitative relations will be presented. For supervised learning I will show how the supervisory burden can be reduced using what we term "deictic supervision", whilst in the unsupervised case I will present a method for learning the most likely interpretation of the training data. I will also show how objects can be "functionally categorised" according to their spatio-temporal behaviour and how the use of type information can help in the learning process, especially in the presence of noise. I will present results from several domains including a kitchen scenario and an aircraft apron.

# Ontology Languages and Engineering

Ian Horrocks

Computing Laboratory  
University of Oxford, UK

Ontologies and ontology based systems are rapidly becoming mainstream technologies, with RDF and OWL now being deployed in diverse application domains, and with major technology vendors starting to augment their existing systems with ontological reasoning. For example, Oracle Inc. recently enhanced its well-known database management system with modules that use RDF/OWL ontologies to support "semantic data management", and their product brochure lists numerous application areas that can benefit from this technology, including Enterprise Information Integration, Knowledge Mining, Finance, Compliance Management and Life Science Research. The design of the high quality ontologies needed to support such applications is, however, still extremely challenging. In this talk I will describe the design of OWL, show how it facilitates the development of ontology engineering tools, describe the increasingly wide range of available tools, and explain how such tools can be used to support the entire design, deployment and maintenance ontology life-cycle.

# Theory of Belief Functions for Data Analysis and Machine Learning Applications: Review and Prospects

Thierry Denoeux

Department of Information Processing Engineering  
Universit de Technologie de Compigne, France

The Dempster-Shafer theory of belief functions provides a unified framework for handling both aleatory uncertainty, arising from statistical variability in populations, and epistemic uncertainty, arising from incompleteness of knowledge. An overview of both the fundamentals and some recent developments in this theory will first be presented. Several applications in data analysis and machine learning will then be reviewed, including learning under partial supervision, multi-label classification, ensemble clustering and the treatment of pairwise comparisons in sensory or preference analysis.

# Modeling Ontological Concepts of Locations with a Heterogeneous Cardinal Direction Model

Hui Shi<sup>1</sup> and Yohei Kurata<sup>2</sup>

<sup>1</sup> SFB/TR8 Spatial Cognition, Universität Bremen, Germany  
Safe and Secure Cognitive Systems, DFKI Bremen, Germany  
shi@informatik.uni-bremen.de

<sup>2</sup> Department of Tourism Science, Tokyo Metropolitan University, Japan  
ykurata@tmu.ac.jp

**Abstract.** Modeling human concepts of object locations is essential for the development of the systems and machines that collaborate with ordinary people on spatial tasks. This paper applies a heterogeneous cardinal direction model, called  $\mathcal{HCDM}$ , to model human concepts of object locations on a plane, using its ability to illustrate where and how an object is located as seen from another with different spatial extensions. For generality, we adopt a set of formal spatial concepts defined in an existing spatial ontology called GUM. These location concepts are associated with the patterns distinguished by  $\mathcal{HCDM}$ . We also discuss the special features of our modeling approach and compare it with the modeling of location concepts by other cardinal direction models.

**Keywords:** Qualitative spatial modeling, cardinal direction models, spatial ontology, object localization.

## 1 Introduction

Modeling how ordinary people conceptualize locations of objects in their living environments is essential for the development of the systems and machines that collaborate with people on spatial tasks, such as ambient assisted living systems, mobile robots, and security monitoring systems, especially if they equip with natural language interfaces. Traditionally, many researchers have discussed a number of expressions and notions that people use for describing object locations, such as *in* (cf. [10]), *left/right of* and *in front of* (cf. [12]), and *over/under* and *above/below* (cf. [5]). This paper adds another speculation on such human concepts of locating objects with the aid of a heterogeneous cardinal direction model, called  $\mathcal{HCDM}$  [11]. This model can be used for capturing where and how an object is located as seen from another object in a 2D space, without regarding to their actual shapes. Accordingly, the model may cover such expressions in natural descriptions as “the potato is *in* a bowl” or “Denmark *borders* Germany *to the north*”.

Typically, in order to describe the location of one object with respect to another, both topological and directional information, often qualitative, are used.

Most qualitative spatial models are classified in *topological* models (cf. Interval Algebra [1], Region Connection calculi [4] and 9-intersection [6]), *directional* models (cf. cardinal direction models [7,13,15], Double-Cross calculus [8], and RfDL [16]), or *combined* models (cf. [18]). Usually, these qualitative models are *homogeneous* in the sense that relevant objects belong to the same domain. However, the cardinal direction model introduced in [11] identifies relations between two objects of different types like points, lines or regions and therefore, this model is *heterogeneous*. Hence we call this model  $\mathcal{HCDM}$  (Heterogeneous Cardinal Direction Model). Furthermore, the model distinguishes the *interior*, *exterior* and *boundary* of both the locatum and relatum and therefore, it is appropriate for modeling both directional and topological relations between objects with different spatial extensions.

The aim of this paper is to analyze the applicability of  $\mathcal{HCDM}$  to model a number of human concepts for locating objects on a plane. Thus, it differs from the work presented in [16], which discusses motion expressions like “go toward”, “pass by” and “across”. Although in natural languages people use thousands of expressions to describe object locations, we consider more generic concepts of locations that underlie such individual expressions. For instance, the expressions “to the left of  $\dots$ ”, as well as “auf der linken Seite  $\dots$ ” in German, are mapped to the same location concept if their slight nuance is ignored. Such generalization is definitely useful to expand the coverage of our approach. For this purpose, we adopt the location concepts defined in an existing spatial ontology, called *Generalized Upper Model extended with space components* (GUM) [3].

This paper is organized as follows: Section 2 reviews the heterogeneous cardinal direction model  $\mathcal{HCDM}$ . Section 3 gives an overview of the ontological concepts specifying locations between objects in GUM. Sections 4 and 5 associate the disjoint and non-disjoint location concepts with spatial patterns modeled by  $\mathcal{HCDM}$ , respectively. The comparison of  $\mathcal{HCDM}$  patterns with spatial relations distinguished by other cardinal direction models are given in Section 6. Finally, Section 7 concludes with a discussion of future work.

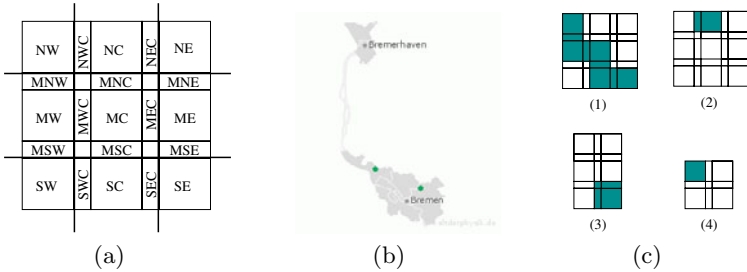
## 2 $\mathcal{HCDM}$ : A Heterogeneous Cardinal Direction Model

$\mathcal{HCDM}$  uses an iconic representation for spatial arrangements of two objects, namely,  $CD^+$ , as an extension of the cardinal direction model [7,9]. It was first introduced in [11]. A  $CD^+$  pattern has  $n \times m$  black-and-white cells, among which the  $i^{th}$  left and  $j^{th}$  bottom cell is marked, if the locatum intersects with the  $i^{th}$  left and  $j^{th}$  bottom field. The fields are determined by four lines,  $l_1 : x = \inf_x(B)$ ,  $l_2 : x = \sup_x(B)$ ,  $l_3 : y = \inf_y(B)$ , and  $l_4 : y = \sup_y(B)$ , which are the greatest lower bound and the least upper bound of the projection of the relatum  $B$ . Accordingly, not only 2D blocks separated by these lines, but also 1D boundaries between two blocks and 0D boundary points between four blocks are all considered independent fields (Fig. 1(a)). This approach is applicable to any pair of objects, but depending on the shape of the relatum some tiles may be collapsed to a line or a point. Naturally,  $5 \times 5$ ,  $3 \times 3$ ,  $5 \times 3$ , and  $3 \times 5$  fields are



distinguished when the relatum is a region or a generic line, a point, a horizontal line, and a vertical line, respectively. In this paper regions refer to simple regions which are single-component regions without disconnected interior, holes, spikes, punctuating points, or cuts.

The  $CD^+$  model is presented in Fig. 1(a), where the names of the one- and two-dimensional fields are given. The four zero-dimensional fields located on the northwest, northeast, southwest and southeast of the middle field are named as  $MNWC$ ,  $MNEC$ ,  $MSWC$  and  $MSEC$ . Fig. 1(b) is the map of the German Federal Land Bremen, including its two cities Bremen and Bremerhaven. The three points represent the three universities: University of Bremen, Jacobs University Bremen and University of Bremerhaven. Fig. 1(c) shows four examples of  $CD^+$  patterns: represents (1) the location of a generic line (the river Weser) as seen from a region (the city Bremen) of the example “the Weser river flows through Bremen from southeast to northwest”; (2) a relation between two regions like “the city Bremerhaven is to the north of the city Bremen”; (3) a relation between a vertical line and a region like “the city Bremen is located on the southeast of the river Lesum”; and (4) a relation between two points like “Jacobs University of Bremen is located to the northwest of the University of Bremen”.



**Fig. 1.** The  $\mathcal{HCDM}$  model and some examples: (a) the model, (b) a map of German Federal Land Bremen, (c) four  $\mathcal{HCDM}$  patterns

Even though  $CD^+$  with  $n \times m$  tiles distinguishes  $2^{n \times m}$  patterns, not all of them are effective as the representation of directional relations between two objects. Some patterns have no geometric instance; for instance, is not allowed because the locatum is a disconnected line or region, and as well, because the locatum as a region has a spike. In [11] a set of constraints are given to exclude geometrically-impossible  $CD^+$  patterns. The  $CD^+$  patterns that satisfy these constraints can be converted to  $CD$  patterns with  $3 \times 3$  black-white cells by integrating the fields on the greatest lower bounds and the least upper bounds of the relatum with their adjacent middle and center fields (as the partition by the thick lines in Fig. 1(a)), in order to enable the unified representation of directional relations by  $CD$  patterns and to reduce the number of spatial arrangements during spatial reasoning. Detailed discussions about the mapping between  $CD^+$  patterns and  $CD$  patterns are given in [11].

### 3 Conceptualization of Object Locations

Recently, based on a series of empirical studies, several ontological representations of natural space and spatial actions have been developed [3]. The intermediate use of such ontological representations allows us to avoid the mapping between countless number of expressions and the domain model. Thus, in this work, we adopt a set of location concepts specified in one of such ontologies, called *Generalized Upper Model extended with space components* (GUM) [3].

To conceptualize the spatial location of an object as seen from a reference object, GUM provides a concept called `SPATIALLOCATING` as a subtype of `Configuration`. A `SpatialLocating` configuration has a placement relation which specifies the place where the entity is being positioned, and the entity itself defined by *locatum* of type `SimpleThing`. The relation `placement` is in fact restricted to the class `GENERALIZEDLOCATION`, which is further specified by a `SpatialModality`. In this paper we treat points, lines and regions as instances of `SimpleThing` and do not specify them further. The following discussion is focused on the classification and specification of `SpatialModality`.

`SpatialModality` has three subtypes: `SpatialDistanceModality` representing spatial distances between objects (e.g., “far/near”); `FunctionalSpatialModality` representing the functional relationship between objects (e.g., “on the table”); and `RelativeSpatialModality` representing topological or projectional relations between objects, which is the major target of this paper. `RelativeSpatialModality` as defined in GUM has a large number of subtypes, among which those relevant to the localization of an object in a 2D space is summarized in Table 1, together with expression examples taken from [3]. To associate these GUM location concepts with *HCDM* patterns (see Section 4 and 5) ontological reasoning is not necessary.

In most cases a *locatum* can be of arbitrary object type (i.e., a point, a line or a region); exceptions are topological relations of type `Peripheral`, `Connection` and `PathRepresenting`. All the concepts in Table 1 can have a region as the *relatum*. A point can be the *relatum* of a disjoint relation of `HorizontalProjectionExternal` or `CardinalDirectionExternal`. Lines can be used as *relatum* of external relations, `PathRepresentingInternal` relations, or `Center` relations.

### 4 Modeling Disjoint Location Concepts by *HCDM*

This and the next sections explore the association of *disjoint* (i.e., the *locatum* and *relatum* have no common point) and *non-disjoint* (i.e., the *locatum* and *relatum* share some point(s)) location concepts with the relation patterns identified by  $CD^+$ , respectively. We consider two scenarios where the *locatum* is a point or a region. On the other hand, the type of the *relatum* is fixed to be simple regions.









As summarized in Table 1 the three external types of spatial concepts `HorizontalProjectionExternal`, `CardinalDirectionExternal` and `PathRepresentingExternal` specify disjoint relations. Although horizontal projections divide the around

**Table 1.** GUM’s specifications relevant to object locations

GUM Specification	Examples	Locatum			Relatum		
		Point	Line	Region	Point	Line	Region
HorizontalProjection External	<i>in front of the car</i> <i>to the left of the sofa</i>	✓	✓	✓	✓	✓	✓
HorizontalProjection Internal	<i>in the front of the car</i>	✓	✓	✓			✓
CardinalDirection External	<i>to the east of Tokyo</i> <i>to the northeast of Tokyo</i>	✓	✓	✓	✓	✓	✓
CardinalDirection Internal	<i>in the east of Tokyo</i>	✓	✓	✓			✓
Center	<i>in the middle of the room</i>	✓	✓	✓		✓	✓
Peripheral	<i>on the edge of the table</i>		✓	✓			✓
Connection	<i>next to the sofa</i>		✓	✓			✓
PathRepresenting External	<i>along the river</i>		✓	✓		✓	✓
PathRepresenting Internal	<i>across the river</i>		✓	✓		✓	✓

space along the two-dimensional axes and cardinal directions use geographical references, the same  $CD^+$  patterns are used to represent corresponding projections and cardinal directions. For instance, the pattern in Fig. 2(a) represents both “to the left front of” and “to the northwest of”. Therefore, in the following discussion we take cardinal direction relations.

#### 4.1 Cardinal Direction Concepts

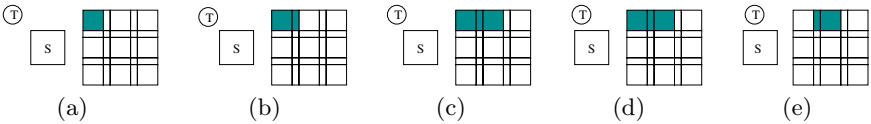
We first consider the cases in which the locatum is a point. The  $CD^+$  model identifies 25 patterns if the relatum is a region and the locatum a point. A point is located to the northwest of a region is represented by the pattern ; to the north by , , , , or possibly , which are mapped to the following two  $CD$  patterns:  and . Since the relatum is not always a rectangle, a point located in the one-dimensional field  $MNC$ , in the zero-dimensional field  $MNWC$  or even in  $MC$ , for example, may represent a north relation depending on the relatum’s shape.

Now we consider the cases in which the locatum is a simple region. This time, however, it is even not always possible to decide whether a pattern fits with a given spatial relation. For instance, the  $CD^+$  pattern in Fig. 2(c) highly fits the relation in “the table is to the northwest of the sofa”, but that in Fig. 2(d) does not so much, although they are represented by the same  $CD^+$  pattern. Obviously, there are a variety of in-between patterns whose characterization is

difficult. On the other hand, we can clearly say that the patterns in Fig. 2(a) and 2(b) fit with the **CardinalDirectionExternal** relation “to the northwest of”, because the most part of the locatum is in the field *NW*. Also, it is clear that the pattern in Fig. 2(e) cannot fit the concept, because no part of locatum is in the field *NW*. From this observation, we have derived the following two conditions:

- $SC_{CD-NW}$  (strong condition of **CardinalDirectionExternal** *northwest*): the spatial relation is mapped to the  $CD^+$  pattern, where the locatum represented by a simple region extends at least the relatum’s *NW*, and may extend *NWC* or *MNW*, but no other fields.
- $WC_{CD-NW}$  (weak condition of **CardinalDirectionExternal** *northwest*): the location pattern is mapped to the  $CD^+$  pattern, where the locatum represented by a simple region extends the relatum’s *NW*, may extend *NWC*, *MNW*, *NC*, *MNC*, *MW*, *MWC*, *MC*, or the zero-dimensional field *MNWC*, but no other fields.

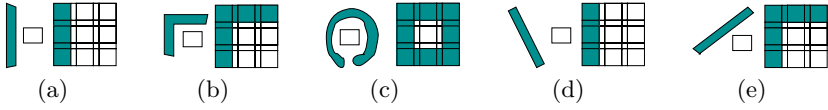
If a  $CD^+$  pattern satisfies the strong condition  $SC_{CD-NW}$ , then the spatial arrangement of two objects always fits with the concept **CardinalDirectionExternal** *northwest* (Figure 2(a) and 2(b)). On the other hand, if a pattern does not satisfy the weak condition  $WC_{CD-NW}$ , then it does not fit with the concept (Figure 2(e)). Lastly, if a pattern satisfies the week condition, but not the strong condition, then it may or may not fit the concept. In this case, we need further criteria to judge whether (or how much) the spatial relation fits with the concept; for instance, the relative area of the region which overlaps with the field *NW*.



**Fig. 2.** (a-c) Location relations that fit with the concept of **CardinalDirectionExternal** “to the northwest of”, but (d) and (e) do not, together with the  $CD^+$  patterns that represent these relations

## 4.2 Path Representing Concepts

Path representing concepts concern the objects both expanding at least one-dimensionally. Therefore, both the locatum and the relatum of a path relation are lines or regions. We consider here the scenario in which both the relatum and the locatum are simple regions. Fig. 3 shows five spatial arrangement, together with the associated  $CD^+$  patterns. Fig. 3(a), 3(b) and 3(c) are relations that fit with the concept **PathRepresentingExternal**, but Fig. 3(d) and 3(e) do not, although the pattern in 3(a) and 3(d), and those in 3(b) and 3(e) are the same.



**Fig. 3.** (a-c) Locations that fit with the concept `PathRepresentingExternal`, but (d) and (e) do not, together with the  $CD^+$  patterns that represent these relations

Here we take the concept `PathRepresentingExternal northwest` (see Fig. 3(b)) as an example and develop its strong and weak conditions. This type of path representing relations depends on the shape of the relatum, which is not considered by the  $CD^+$  model. Even if the locatum extends completely in the west and north sides of the relatum, the locatum is not necessarily located around the relatum to its northwest (compare Fig. 3(b) and 3(e)). Therefore, the strong condition for this concept does not exist. On the other hand, if the locatum extends in the west and north sides of the relatum, but no other side, the locatum may be located around the relatum to its northwest. Thus, we give the following definition of the weak condition.

- $WC_{CD-NW}$  (weak condition of `PathRepresentingExternal northwest`): the location pattern is mapped to  $r : l$ , where the locatum  $l$  represented by a simple region extends at least the relatum  $r$ 's  $WM, MNW, NW, NWC$  and  $NC$ , but neither  $SEC, SE$  and  $MSE$ , nor  $MSEC$ .

The developed conditions for disjoint spatial relations between simple regions are summarized in Table 2. The conditions are visualized by icons with  $5 \times 5$  cells, which are marked by three colors: black, gray, and white cells indicate the fields over which the region must, may, and cannot extend. Under each icon are the numbers of  $CD^+$  and  $CD$  patterns that satisfy the corresponding condition, respectively.

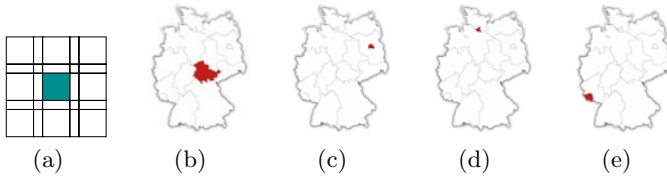
**Table 2.** Strong and weak conditions of disjoint location concepts, together with the numbers of the  $CD^+$  and  $CD$  patterns that satisfy each condition

	CardinalDirectionExternal		PathRepresentingExternal		
	<i>north</i>	<i>northwest</i>	<i>north</i>	<i>northwest</i>	<i>surrounding</i>
Strong Condition					
	4, 1	4, 4			1, 1
Weak Condition					
	1303, 25	53, 7	1141, 25	2772, 24	496, 2

## 5 Modeling Non-disjoint Location Concepts by $\mathcal{HCDM}$

### 5.1 Parthood

In parthood concept the locatum is contained by the relatum. These relations are specified by the GUM concepts `HorizontalProjectionInternal` (e.g., “in the front of”) and `CardinalDirectionInternal` (e.g., “in the north of”), as well as `Center` (e.g., “in the center of”). In a parthood pattern, the locatum extends only the field  $MC$  and its adjacent zero- and one-dimensional fields. For example, the only  $CD^+$  pattern that satisfies `Center` is Fig. 4(a); the same pattern may also represent all possible internal cardinal direction or horizontal projection relations, as depicted in Fig. 4, where Fig. 4(b) shows “Thüringen is in the middle of Germany”, 4(c) “Berlin is in the northeast of Germany”, 4(d) “Hamburg is in the north of Germany”, and 4(e) “Saarland is in the southwest of Germany and borders to French”. Obviously,  $CD^+$  (or  $CD$ ) patterns are insufficient for representing internal cardinal direction relations, since  $CD^+$  cannot distinguish a variety of parthood patterns. It is also the case if the objects to be located are points or lines.

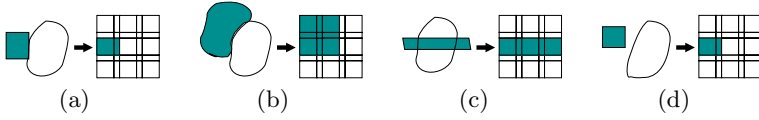


**Fig. 4.** (a)  $CD^+$  pattern represents the cardinal direction relations (b-e); (b) “in the center of”, (c) “in the northeast of”, (d) “in the north of” and (e) “in the southeast of”.

### 5.2 Overlapping and Connection

The connective relations can be further divided by cardinal directions or horizontal projections, e.g., “border to  $\dots$  on the north” or “in the left end of”. Similarly, the concept `PathRepresentingInternal` may contain cardinal direction or horizontal projection information, e.g., “across the park from west to east”. Fig. 5 shows some connection and overlapping patterns, together with their associated  $CD^+$  patterns. Fig. 5(a) may model “to border  $\dots$  to the west”, 5(b) “to border  $\dots$  to the northwest”, and 5(c) “to across  $\dots$  from west to east”, but the spatial relation in 5(d) does not fit `Connection` concept, though it is represented by the same pattern of the relation in 5(a).

Comparing 5(a) and 5(d), we can see that a given  $CD^+$  pattern may or may not be associated with connection or overlapping concepts. since the connection concept depends not only on the cardinal direction or horizontal projection information about where they are connected, but also on the shapes of both objects, whereas  $CD^+$  does not capture any shape information. Accordingly, there are no



**Fig. 5.** Four spatial relations and their  $CD^+$  pattern representations: (a) and (b) fit Connection concept; (c) fits PathRepresentingInternal concept; but (d) does not fit Connection concept

strong conditions for connection relations. The following is the weak condition for the concept Connection on *west* as an example.

- $WC_{CD-NW}$  (weak condition of Connection *west*): the location pattern is mapped to  $r : l$ , where the locatum  $l$  represented by a simple region extends at least the relatum  $r$ 's *MW* and *MWC*.

Table 3 summarizes the strong and weak conditions of the GUM location concepts regarding connection and overlapping, where “-” means that almost all relations satisfy the condition and accordingly, they are not interesting for any computational purposes.

**Table 3.** Strong and weak conditions of Connection and PathRepresentingInternal, together with the numbers of  $CD^+$  and  $CD$  patterns that satisfy each condition

	Connection		PathRepresentingInternal		
	<i>west</i>	<i>northwest</i>	<i>west-east</i>	<i>west-north</i>	<i>southwest-northeast</i>
Strong Condition					
	0, 0	0, 0	625, 1	520, 1	169, 9
Weak Condition					
	-, 100	-, 28	-, 64	-, 56	-, 32

### 5.3 Discussions

Comparing the modeling results presented in Tables 2 and 3, it is clear that both  $CD^+$  and  $CD$  are suitable for representing location concepts specifying **disjoint** spatial arrangements, especially, disjoint cardinal direction and horizontal projection concepts. In non-disjoint cases, the numbers of  $CD^+$  patterns that satisfy the weak conditions are very large and therefore, they are not useful for practical applications. Moreover,  $CD^+$  (or  $CD$ ) cannot distinguish connection relations between two simple regions, since whether two regions are connected depends on their shapes, which cannot be modeled by  $CD^+$ . However, the strong conditions

for deciding how a path represented as a region overlaps with a reference region are useful.

Although  $CD^+$  distinguishes much more spatial relations, in most cases it is too complex for computational processes.  $CD^+$  patterns are mapped to  $CD$  patterns, the numbers of patterns which satisfy each concept are reduced significantly. Thus, it is better to primarily use lower-grained  $CD$  patterns for the spatial inference.

## 6 Comparison with Other Cardinal Direction Models

*Rectangle Algebra*, as a natural generalization of *Interval Algebra* [1] to the two-dimensional space, is introduced in [2]. This model targets rectangles, whose sides are parallel to the axes of a 2D space. A relation between two rectangles is represented as a pair  $(R_x, R_y)$  of interval relations: the x-axis and y-axis relations. There are  $13 \times 13 = 169$  possible basic relations between any two given rectangles. If the connectivity in interval relations is ignored, there remains six basic interval relations, and  $6 \times 6 = 36$  basic rectangle relations, which are captured by *Rectangle Cardinal Direction* model [17]. The model captures how the locatum intersects with the  $3 \times 3$   $CD$  tiles. If the domain is restricted to rectangles, the  $CD^+$  model is equivalent to the *Rectangle Algebra*, and  $CD$  to the *Rectangle Cardinal Direction* model. On the other hand, the existing cardinal direction calculi for the relations between points (cf. [7,13]) can be seen as a spatial case of the  $\mathcal{HCDM}$  model, where the relatum is represented as a point (see the pattern (4) in Fig. 1(c)).

[14] introduces *Symbolic Spatial Indexes* for representing symbolic images or geographical maps in a 2D space, where every spatial relation is defined between the direction and topological representative points of related objects. To capture the direction of an object, a number of representative points, usually centers and edge points, of the relatum and locatum will be taken. These points partition the space around the relatum into a set of zero-, one- and two-dimensional fields. Similarly, topological representative points are identified to define topological relations using symbolic spatial indexes. Therefore, the spatial relations between two region objects depend strongly on the selection of representative points, different representative points may lead to different spatial relations between these objects.

The model presented in [9] uses a  $3 \times 3$  matrix, called *deep direction-relation matrix*, to represent direction relations between two objects, in order to use them in spatial database query for geographic information systems. This model is also a heterogeneous model and models relations between points, lines and regions. However, since a  $3 \times 3$  matrix provides no account for the boundary information of the relatum, the *neighbor code* of each matrix element is used to encode such information, which is not intuitive as  $\mathcal{HCDM}$ .

From the above analysis, we consider that  $\mathcal{HCDM}$  is a highly generic model for representing location concepts between different types of objects. Generally, the  $\mathcal{HCDM}$  model distinguishes a larger number of spatial patterns (Table 2 and



3). Accordingly, the computational complexity of operations on these patterns is high. In [11] we have defined the general *inverse* and *composition* operations on  $\mathcal{HCDM}$  patterns.  $\mathcal{HCDM}$  is reduced to *Rectangle Algebra* [2], if objects are rectangles. Similarly,  $\mathcal{HCDM}$  is reduced to *Rectangle Cardinal Relations* model [17] if the topological relations between objects can be ignored; and to point based cardinal direction calculi [7,13] for modeling relations between points. Therefore, spatial inference methods developed for those models and calculi can be reused to reasoning about  $\mathcal{HCDM}$  pattern in each situation.

## 7 Conclusion

Modeling human location concepts is necessary for enriching the communication between people and computers/machines collaborating on spatial tasks. This paper explored the modeling of a variety of relative spatial arrangements of an object with respect to another one. To represent such arrangements between various types of objects, the heterogenous cardinal direction model  $\mathcal{HCDM}$  was applied. In order to characterize some arrangements, we may need further criteria other than  $\mathcal{HCDM}$  patterns, which are left for future work. This paper also demonstrated that the specification in GUM, as an upper model, is very useful to capture a number of location concepts in a generic and domain-independent way.

We are currently investigating the application of our findings to the interface of an ambient assisted living environment, such that elderly or impaired people can intuitively interact with the environment by localizing objects using natural language. To achieve this goal, reasoning algorithms are necessary for supporting efficient localization processes.

**Acknowledgments.** We gratefully acknowledge the support of the Deutsche Forschungsgemeinschaft (DFG) through the Collaborative Research Center SFB/TR 8 Spatial Cognition - Subproject I3-SharC.

## References

1. Allen, J.F.: Maintaining knowledge about temporal intervals. *CACM* 26(11), 832–843 (1983)
2. Balbiani, P., Condotta, J., del Cerro, L.F.: A model for reasoning about bidimensional temporal relations. In: *Proc. of the Sixth International Conference On Principles of Knowledge Representation and Reasoning*, pp. 124–130 (1998)
3. Bateman, J.A., Hois, J., Ross, R.J., Tenbrink, T.: A Linguistic Ontology of Space for Natural Language Processing. *Artificial Intelligence* (accepted)
4. Cohn, A.G., Bennett, B., Gooday, J., Gotts, N.M.: Qualitative spatial representation and reasoning with the region connection calculus. *Geoinformatics* 1, 1–44 (1997)
5. Coventry, K.R., Prat-Sala, M., Richards, V.L.: The interplay between geometry and function in the comprehension of "over", "under", "above" and "below". *Journal of Memory and Language* 44, 376–398 (2001)

6. Egenhofer, M., Mark, D.: Modeling conceptual neighborhoods of topological line-region relations. *International Journal of Geographical Information Systems* 9(5), 555–565 (1995)
7. Frank, A.: Qualitative spatial reasoning about cardinal directions. In: *ACSM-ASPRS Auto-Carto 10* (1991)
8. Freksa, C.: Using orientation information for qualitative spatial reasoning. In: Frank, A.U., Formentini, U., Campari, I. (eds.) *GIS 1992*. LNCS, vol. 639, pp. 162–178. Springer, Heidelberg (1992)
9. Goyal, R., Egenhofer, M.: Consistent queries over cardinal directions across different levels of detail. In: *11th International Workshop on Database and Expert Systems Applications*, pp. 876–880 (2000)
10. Herskovits, A.: *Language and Spatial Cognition: An Interdisciplinary Study of the Prepositions in English*. Cambridge University Press, Cambridge (1986)
11. Kurata, Y., Shi, H.: Toward heterogeneous cardinal direction calculus. In: Mertsching, B., Hund, M., Aziz, Z. (eds.) *KI 2009*. LNCS (LNAI), vol. 5803, pp. 452–459. Springer, Heidelberg (2009)
12. Levinson, S.C.: *Space in Language and Cognition: Explorations in Cognitive Diversity*. Cambridge University Press, Cambridge (2003)
13. Ligozat, G.: Reasoning about cardinal directions. *Journal of Visual Languages and Computing* 9, 23–44 (1998)
14. Papadias, D., Sellis, T.: Qualitative representation of spatial knowledge in two dimensional space. *Journal of Very Large Data Bases. Special Issue on Spatial Databases* 3(4), 479–516 (1994)
15. Renz, J., Mitra, D.: Qualitative direction calculi with arbitrary granularity. In: Zhang, C., Guesgen, H.W., Yeap, W.-K. (eds.) *PRICAI 2004*. LNCS (LNAI), vol. 3157, pp. 65–74. Springer, Heidelberg (2004)
16. Shi, H., Kurata, Y.: Modeling ontological concepts of motions with two projection-based spatial models. In: *2nd International Workshop on Behavioral Monitoring and Interpretation (in conjunction with KI-2008)*, pp. 42–56 (2008)
17. Skiadopoulos, S., Koubarakis, M.: Composing cardinal direction relations. *Artificial Intelligence* 152(2), 147–171 (2004)
18. Sun, H., Li, W.: Integrated qualitative spatial reasoning. In: *Proc. of the 2004 International Conference on Computational Intelligence*, pp. 341–344 (2004)

# Finding Minimal Rare Itemsets and Rare Association Rules

Laszlo Szathmary<sup>1</sup>, Petko Valtchev<sup>1</sup>, and Amedeo Napoli<sup>2</sup>

<sup>1</sup> Dépt. d'Informatique UQAM, C.P. 8888,  
Succ. Centre-Ville, Montréal H3C 3P8, Canada

Szathmary.L@gmail.com, valtchev.petko@uqam.ca

<sup>2</sup> LORIA UMR 7503, B.P. 239, 54506 Vandœuvre-lès-Nancy Cedex, France  
napoli@loria.fr

**Abstract.** Rare association rules correspond to rare, or infrequent, itemsets, as opposed to frequent ones that are targeted by conventional pattern miners. Rare rules reflect regularities of local, rather than global, scope that can nevertheless provide valuable insights to an expert, especially in areas such as genetics and medical diagnosis where some specific deviations/illnesses occur only in a small number of cases. The work presented here is motivated by the long-standing open question of efficiently mining strong rare rules, i.e., rules with high confidence and low support.

## 1 Introduction

Conventional pattern miners target the frequent itemsets and rules in a dataset. These are believed to reflect the globally valid trends and regularities dug in the data, hence they typically support modelling and/or prediction. Yet in many cases global trends are known or predictable beforehand by domain experts, therefore such patterns do not bear much value to them. In contrast, regularities of local scope, i.e., covering only a small number of data records, or transactions, may be of higher interest as they could translate less well-known phenomena, e.g., contradictions to the general beliefs in the domain or notable exceptions thereof [1]. This is often true in areas such as genetics and medical diagnosis where many deviations / symptom combinations will only manifest in a small number of patient cases. Hence the potential of the methods for mining the corresponding patterns and rules for supporting a more focused analysis of the recorded biomedical data.

### 1.1 Motivating Examples

A first case study for atypical patterns and rules pertains to a French biomedical database, the STANISLAS cohort [2]. The STANISLAS cohort comprises the medical records of a thousand presumably healthy French families. In a particular problem settings, the medical experts are interested in characteristics and relations that pertain to a very small number of individuals. For instance, a key

goal in this context is to investigate the impact of genetic and environmental factors on diversity in cardiovascular risk factors. Interesting information to extract from the cohort database includes the patient profiles associating genetic data with extreme or borderline values of biological parameters. However, such types of associations should be atypical in healthy cohorts.

To illustrate the concept of rare rules and its potential benefits, assume we want to target the causes for a group of cardiovascular diseases (CVD) within the STANISLAS cohort. If a frequent combination of CVD and a potential factor is found, then the factor may be reasonably qualified as a facilitator for the disease. For instance, a frequent itemset “{elevated cholesterol level, CVD}” and a strong association rule “{elevated cholesterol level} $\Rightarrow$  {CVD}” would empirically validate the widely acknowledged hypothesis that people with high cholesterol level are at serious risk of developing a CVD. In contrast, if the itemset involving a factor and CVD is rare, this would suggest an inhibiting effect on the disease. For instance, the rareness of the itemset “{vegetarian, CVD}” would suggest that a good way to reduce the CVD risk is to observe a vegetarian diet.

The second case study pertains to pharmacovigilance, a domain of pharmacology dedicated to the detection, monitoring and study of adverse drug effects. Given a database of clinical records together with taken drugs and adverse effects, mining relevant itemsets would enable a formal association between drugs adverse effects. Thus, the detected patterns of (combinations of) drugs with undesired (or even fatal) effects on patients could provide the basis for an informed decision as to the withdrawal or continuance of a given drug. Such decision may affect specific patients, part of or event in the entire drug market (see, for instance, the withdrawal of the lipid-lowering drug *Cerivastatin* in August 2001). Yet in order to make appear the alarming patterns of adverse effects, the benign ones, which compose the bulk of the database content, should be filtered out first. Once again, there is a need to skip the typical phenomena and to focus on less expectable ones. It is noteworthy that similar reasoning may be abstracted from unrelated problem domains such as bank fraud detection where fraudulent behaviour patterns manifest in only a tiny portion of the transaction database content.

## 1.2 Approaches and Recent Progress

Pattern mining based on the support metrics is biased upon the detection of trends that are – up to a tolerance threshold – globally valid. Hence a straightforward approach to the detection of atypical and local regularities has been to relax the crisp and uniform minimal support criterion for patterns [3].

In a naïve problem settings, the minimal support could be decreased sufficiently to include in the frequent part of the pattern family all potentially interesting regularities. Yet this would have a devastating impact on the performances of the pattern miner on top of the additional difficulties in spotting the really interesting patterns within the resulting huge output (known as the *rare item problem* [4,5]).

A less uniform support criterion is designed in [5] where the proposed method *RSAA* (Relative Support Apriori Algorithm) relies on item-wise minimal support thresholds with user-provided values. *RSAA* outputs all itemsets, and hence rules, having their support above at least one support threshold corresponding to a member item. Thus, the output still comprises all frequent itemsets and rules together with some, but not necessarily all, atypical ones.

A higher degree of automation is achieved in *MSapriori* (Multiple Supports Apriori) [4] by modulating the support of an itemset with the supports of its member items. Thus, the support is increased by a factor inversely proportional to the lowest member support, which, on the bottom line increases the chances of itemsets involving infrequent items to nevertheless make it to the frequent part of the pattern family. Once more, the overall effect is the extension of the frequent part in the pattern family by some infrequent itemsets.

Our own approach is a more radical departure from the standard pattern mining settings as it focuses directly on the infrequent part of the pattern family that becomes the mining target. The underlying key notion is the *rare itemset (rule)* defined as an itemset (rule) with support lower than the threshold. *Apriori-Inverse* [6], and *MIISR* (Mining Interesting Imperfectly Sporadic Rules) [7] are two methods from the literature that exploit the same rarity notion, yet the former would exclusively mine perfectly rare itemsets (i.e., having exclusively rare subsets) while the latter slightly relaxes this overtly crisp constraint. This, on the bottom line, amounts to exploring rare patterns within the order filter above the rare singleton itemsets (i.e., rare items) in the itemset lattice while ignoring rare itemsets mixing both rare and frequent items.

In our own approach, we concentrate on the dual part of the frequent subfamily, i.e., on all rare itemsets and not merely the perfectly rare ones. To that end, we devised a strategy that traverses the frequent zone of the itemset lattice (the order ideal of the frequent itemsets) at minimal cost, as described in [8]. The current paper is a follow-up dealing with rare rule generation out of the resulting set of rare itemsets (see next section).

It is noteworthy that playing with minimal support is not the only way to approach the mining of atypical regularities. Thus, different statistical measures may be used to assess atypicality of patterns that are not bound to the number of occurrences. Moreover, the availability of an explicitly expressed body of expert knowledge or expectations/beliefs (e.g., as general rules) for a particular dataset or analysis problem enables a more focused pattern extraction where an unexpected or exceptional pattern is assessed with respect to a generally admitted one (a relevant discussion thereof may be found in [9]).

Rare itemsets, similarly to frequent ones, could be easily turned into rules, i.e. by splitting them into premise and conclusion subsets. The resulting rules are necessarily rare but their confidence would vary. Only rules of high confidence can be reasonably considered as regularities.

The extraction of rare itemsets and rules presents significant challenges for data mining algorithms [3]. In particular, algorithms designed for frequent itemset mining are inadequate for extracting rare association rules. Therefore, new

specific algorithms have to be designed. The problem with conventional frequent itemset mining approaches is that they have a (physical) limit on how low the minimum support can be set. We call this absolute limit the *barrier*: the barrier is the absolute minimum support value that is still manageable for a given frequent itemset mining algorithm in a given computing environment. The exact position (value) of the barrier depends on several variables, such as: (1) the database (size, density, highly- or weakly-correlated, etc.); (2) the platform (characteristics of the machine that is used for the calculation (CPU, RAM)); (3) the software (efficient / less efficient implementation), etc. Conventional search techniques are *always* dependent on a physical limit that cannot be crossed: it is almost certain that the minimum support cannot be lowered to 1.<sup>1</sup> The questions that arise are: how can the barrier be crossed; what is on the other side of the barrier; what kind of information is hidden; and mainly, how to extract interesting association rules from the negative side of the barrier.

### 1.3 Contribution

In order to generate rare association rules, first rare itemsets have to be extracted. In [10] it is stated that the negative border of frequent itemsets can be found with levelwise algorithms. A straightforward modification of the *Apriori* algorithm has been proposed in [8] for this task. During the levelwise search, *Apriori* computes the support of *minimal rare itemsets* (mRIs), i.e. rare itemsets such that all proper subsets are frequent. Instead of pruning the mRIs, they are retained. In addition, it is shown that the mRIs form a generator set of rare itemsets, i.e. *all rare itemsets* can be restored from the set of mRIs [8].

In this paper, we focus on the search for valid rare association rules, i.e. rules with low support and high confidence. Once all rare itemsets are available, in theory it is possible to generate all valid rare association rules. However, this method has two drawbacks. First, the restoration of all rare itemsets is a very memory-expensive operation due to the huge number of rare itemsets. Second, having restored all rare itemsets, the number of generated rules would be even more. Thus, the same problem as in the case of frequent valid association rules has to be faced: dealing with a huge number of rules of which many are redundant and not interesting at all.

Frequent itemsets have several condensed representations, e.g. closed itemsets, generators representation, free-sets, non-derivable itemsets, etc. However, from the application point of view, the most useful representations are closed itemsets and generators. Among frequent association rules, bases are special rule subsets from which all other frequent association rules can be restored with a proper inference mechanism. The set of minimal non-redundant association rules ( $\mathcal{MNR}$ ) is particularly interesting, because it is a lossless, sound, and informative representation of all valid (frequent) association rules [11]. Moreover, these frequent rules allow one to deduce a maximum of information with minimal

---

<sup>1</sup> When the absolute value of minimum support is 1, then all existing itemsets are frequent.

hypotheses. Accordingly, the same sort of subset has been searched for rare rules, namely the set of minimal rare itemset rules, presented hereafter.

The present work is motivated by the long-standing open question of devising an efficient algorithm for finding rules that have a high confidence together with a low support. This work shows a number of characteristics that are of importance. First, valid rare association rules can be extracted efficiently. Second, an interesting subset of rare association rules can be directly computed, similar to the set of (frequent)  $\mathcal{MNR}$  rules in the case of frequent rules. Third, the method is rather easy to implement.

The paper is organized as follows. The basic concepts and definitions for frequent and rare itemsets are presented in Section 2. Then, Section 3 details the generation of informative rare association rules from rare itemsets. Finally, Section 4 concludes the paper.

## 2 Frequent and Rare Itemsets

Consider the following  $5 \times 5$  sample dataset:  $\mathcal{D} = \{(1, ABDE), (2, AC), (3, ABCE), (4, BCE), (5, ABCE)\}$ . Throughout the paper, we will refer to this example as “**dataset  $\mathcal{D}$** ”.

We consider a set of *objects* or *transactions*  $\mathcal{O} = \{o_1, o_2, \dots, o_m\}$ , a set of *attributes* or *items*  $\mathcal{A} = \{a_1, a_2, \dots, a_n\}$ , and a relation  $\mathcal{R} \subseteq \mathcal{O} \times \mathcal{A}$ . A set of items is called an *itemset*. Each transaction has a unique identifier (*tid*), and a set of transactions is called a *tidset*. The tidset of all transactions sharing a given itemset  $X$  is its *image*, denoted  $t(X)$ . For instance, the image of  $\{A, B\}$  in  $\mathcal{D}$  is  $\{1, 3, 5\}$ , i.e.,  $t(AB) = 135$  in our separator-free set notation. The *length* of an itemset  $X$  is  $|X|$ , whereas an itemset of length  $i$  is called an  $i$ -itemset. The (absolute) *support* of an itemset  $X$ , denoted by  $\text{supp}(X)$ , is the size of its image, i.e.  $\text{supp}(X) = |t(X)|$ . Moreover,  $X$  is *frequent*, if its support is not less than a given *minimum support* threshold  $\text{min\_supp}$ , i.e.  $\text{supp}(X) \geq \text{min\_supp}$ . Dually, if a maximal support threshold  $\text{max\_supp}$  is provided then an itemset  $P$  such that  $\text{supp}(P) \leq \text{max\_supp}$  is called *rare* (or *infrequent*). If the support of an itemset is 0 then the itemset is a *zero itemset*<sup>2</sup>, otherwise it is a *non-zero itemset*.

An equivalence relation is induced by  $t$  on the power-set of items  $\wp(\mathcal{A})$ : equivalent itemsets share the same image ( $X \cong Z$  iff  $t(X) = t(Z)$ ) [12]. Consider the equivalence class of  $X$ , denoted  $[X]$ , and its extremal elements w.r.t. set inclusion.  $[X]$  has a unique maximum (a *closed* itemset), and a set of minima (*generator* itemsets). A *singleton* equivalence class has only one element. The following definition exploits the monotony of support upon set inclusion in  $\wp(\mathcal{A})$ :

**Definition 1.** *An itemset  $X$  is closed (generator) if it has no proper superset (subset) with the same support.*

---

<sup>2</sup> Not to be confused with the empty set.

A *closure* operator underlies the set of closed itemsets; it assigns to  $X$  the maximum of  $[X]$  (denoted by  $\gamma(X)$ ). Naturally,  $X = \gamma(X)$  for closed  $X$ . Generators, a.k.a. *key-sets* in database theory, represent a special case of free-sets [13]. The following property, which is widely known in the domain, basically states that the generator family is a downset within the Boolean lattice  $\langle \wp(\mathcal{A}), \subseteq \rangle$ :

*Property 1.* Given  $X \subseteq \mathcal{A}$ , if  $X$  is a generator, then  $\forall Y \subseteq X$ ,  $Y$  is a generator, whereas if  $X$  is not a generator,  $\forall Z \supseteq X$ ,  $Z$  is not a generator.

The separation of  $\wp(\mathcal{A})$  into rare and frequent parts induces substructures that reflect the same extremum principle as generators/closures but in a global scope rather than within a single equivalence class.

**Definition 2.** (i) A frequent itemset is a maximal frequent itemset (MFI) if all its proper supersets are not frequent. (ii) An itemset is a minimal rare itemset (mRI) if it is rare, and all its proper subsets are not rare. (iii) A minimal rare generator (mRG) is a rare generator such that and all its proper subsets are not rare.

In [8] we showed that mRIs are in fact generators, i.e. the set of mRIs and the set of mRGs are equivalent:

**Proposition 1.** All minimal rare itemsets are generators [8].

In the general problem settings, an interval may exist between the thresholds  $min\_supp$  and  $max\_supp$ . Yet throughout the paper we shall assume that both values describe a unique separation of  $\wp(\mathcal{A})$  into a frequent and a rare part (i.e. there will be no itemsets that are neither rare nor frequent). This basically means, in absolute terms, that  $max\_supp = min\_supp - 1$ .

The above equality amounts to the existence of cut across the powerset lattice separating the frequent part from the rare one. This cut, called hereafter the *border* as in [10], has a positive side, made of the frequent itemsets, and a negative side, made of the rare itemsets. Both sides of the border have intriguing mathematical properties (see [13,14]) whereas their computation has been reduced to well-known combinatorial generation problems [15].

## 3 Rare Association Rules

### 3.1 Basic Concepts

An association rule is an expression of the form  $P_1 \rightarrow P_2$ , where  $P_1$  and  $P_2$  are arbitrary itemsets ( $P_1, P_2 \subseteq \mathcal{A}$ ),  $P_1 \cap P_2 = \emptyset$  and  $P_2 \neq \emptyset$ . The left side,  $P_1$  is called *antecedent*, the right side,  $P_2$  is called *consequent*. The support of an association rule  $r: P_1 \rightarrow P_2$  is defined as:  $supp(r) = supp(P_1 \cup P_2)$ . The *confidence* of an association rule  $r: P_1 \rightarrow P_2$  is defined as the conditional probability that an object includes  $P_2$ , given that it includes  $P_1$ :  $conf(r) = supp(P_1 \cup P_2) / supp(P_1)$ . An association rule  $r$  is called *confident*, if its confidence is not less than a given *minimum confidence* (denoted by  $min\_conf$ ), i.e.  $conf(r) \geq min\_conf$ . An



association rule  $r$  with  $\text{conf}(r) = 1.0$  (i.e. 100%) is an *exact* association rule, otherwise it is an *approximate* association rule.

An association rule  $r$  is called *frequent* if its support is not less than a given *minimum support* (denoted by  $\text{min\_supp}$ ), i.e.  $\text{supp}(r) \geq \text{min\_supp}$ . A frequent association rule is *valid* if it is confident, i.e.  $\text{supp}(r) \geq \text{min\_supp}$  and  $\text{conf}(r) \geq \text{min\_conf}$ . *Minimal non-redundant association rules* ( $\text{MNR}$ ) have the following form:  $P \rightarrow Q \setminus P$ , where  $P \subset Q$  and  $P$  is a frequent *generator* and  $Q$  is a frequent *closed* itemset.

An association rule is called *rare* if its support is not more than a given *maximum support*. Since we use a single border, it means that a rule is rare if its support is less than a given *minimum support*. A rare association rule  $r$  is *valid* if  $r$  is confident, i.e.  $\text{supp}(r) < \text{min\_supp}$  and  $\text{conf}(r) \geq \text{min\_conf}$ . In the rest of the paper, by “rare association rules” we mean *valid* rare association rules.

### 3.2 Breaking the Barrier

Recall that our goal is to break the *barrier*, i.e. to be able to extract rare itemsets and rare association rules that cannot be extracted with the direct approach used by conventional frequent itemset mining algorithms like *Apriori*. With the *BtB* (Breaking the Barrier) algorithm we can extract highly confident rare association rules below the barrier. The algorithm consists of the following three main steps.

*First*, for computing the set of minimal rare itemsets, the key algorithm is *Apriori-Rare* [8]. *Apriori* finds frequent itemsets, but as a “side effect” it also explores the so-called minimal rare itemsets (mRIs). *Apriori-Rare* retains these itemsets instead of pruning them. In Section 2 we show that minimal rare itemsets are rare generators (see Proposition 1).

*Second*, find the closures of the previously found minimal rare itemsets so as to obtain their equivalence classes.

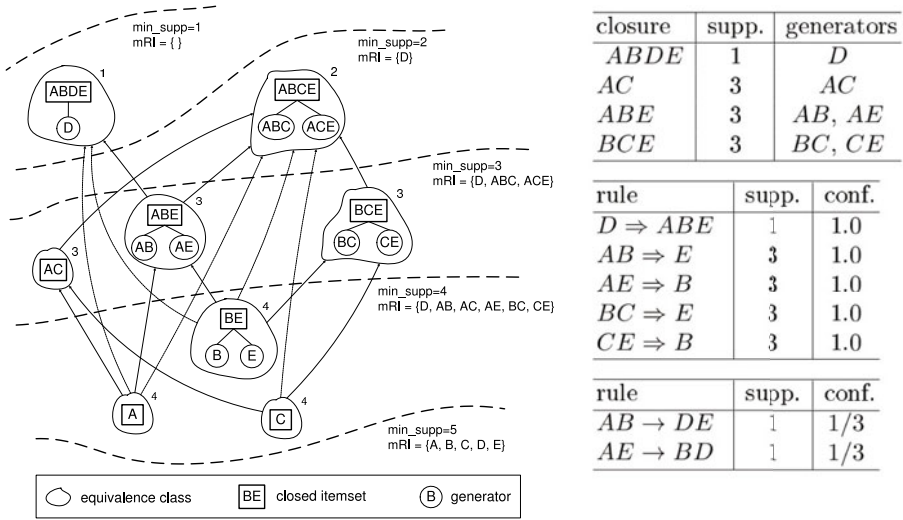
*Third*, from the explored rare equivalence classes it is possible to generate rare association rules in a way very similar to that of finding (frequent) minimal non-redundant association rules. We call these rare rules “mRG rules” because their antecedents are minimal rare generators.

### 3.3 mRG Rules

Two kinds of mRG rules can be distinguished, namely exact and approximate rules. In this paper we concentrate on exact mRG rules that can be characterized as:

$$r: P_1 \Rightarrow P_2 \setminus P_1, \text{ where } \begin{array}{l} P_1 \subset P_2 \\ P_1 \text{ is an mRG} \\ P_1 \cup (P_2 \setminus P_1) = P_2 \text{ is a rare closed itemset} \\ \text{conf}(r) = 1.0 \end{array}$$

From the form of exact mRG rules it follows that these rules are *rare* association rules, where the antecedent ( $P_1$ ) is rare and the consequent ( $P_2 \setminus P_1$ ) is rare *or* frequent.  $P_1$  and  $P_2$  are in the same equivalence class.



**Fig. 1. Left:** rare equivalence classes found by *BtB* in dataset  $\mathcal{D}$  at different *min\_supp* values. **Top right:** rare equivalence classes found by *BtB* in  $\mathcal{D}$  with *min\_supp* = 4. **Center right:** exact mRG rules in  $\mathcal{D}$  with *min\_supp* = 4. **Bottom right:** approximate mRG rules in  $\mathcal{D}$  with *min\_supp* = 4.

Since a generator is a minimal subset of its closure with the same support, these rules allow us to deduce maximum information with minimal hypothesis, just as the  $\mathcal{MNR}$  rules. Using Kryszkiewicz’s cover operator [16], one can restore further *exact* rare association rules from the set of exact mRG rules.

*Example.* Figure 1 (left) shows all the equivalence classes of dataset  $\mathcal{D}$ . Support values are depicted above to the right of equivalence classes. Itemsets with the same support are grouped together in the same level. Levels are separated by borders that are defined by different *min\_supp* values. Next to each *min\_supp* value, the corresponding minimal rare itemsets are also shown. For instance, if *min\_supp* = 4 then there exist 5 frequent itemsets ( $A, C, B, E, BE$ ) and 6 minimal rare itemsets ( $D, AB, AC, AE, BC, CE$ ).

Suppose that the barrier is at *min\_supp* = 4. In this case, using *Apriori*, the less frequent association rules have support 4. With *Apriori-Rare*, the following mRIs are found:  $D, AB, AC, AE, BC$  and  $CE$ . Calculating their closures, four rare equivalence classes are explored, as shown in Figure 1 (top right). Note that *not all* rare equivalence classes are found. For instance, the class whose maximal element is  $ABCE$  is not found because its generators are *not* mRIs, i.e. it is not true for  $ABC$  and  $ACE$  that all their proper subsets are frequent itemsets.

*Generating exact mRG rules.* Once rare equivalence classes are found, the rule generation method is basically the same as in the case of  $\mathcal{MNR}$  rules. Exact mRG rules are extracted within the same equivalence class. Such rules can only

be extracted from non-singleton classes. Figure 1 (center right) shows which exact mRG rules can be extracted from the found rare equivalence classes (Figure 1, top right).

*Generating approximate mRG rules.* Approximate mRG rules are extracted from classes whose maximal elements are comparable with respect to set inclusion. Let  $P_1$  be an mRG,  $\gamma(P_1)$  the closure of  $P_1$ , and  $[P_1]$  the equivalence class of  $P_1$ . If a proper superset  $P_2$  of  $\gamma(P_1)$  is picked among the maximal elements of the found rare equivalence classes different from  $[P_1]$ , then  $P_1 \rightarrow P_2 \setminus P_1$  is an approximate mRG rule. Figure 1 (bottom right) shows the approximate mRG rules that can be extracted from the found rare equivalence classes (Figure 1, top right).

### 3.4 Experimental Results

In this section we present the results of a series of tests. First, we provide results that we obtained on a real-life biomedical dataset. Then, we demonstrate that our approach is computationally efficient for extracting rare itemsets and rare association rules. Thus, a series of computational times resulting from the application of our algorithms to well-known datasets is presented. All the experiments were carried out on an Intel Pentium IV 2.4 GHz machine running under Debian GNU/Linux operating system with 512 MB RAM. Algorithms were implemented in the CORON platform [17].<sup>3</sup> All times reported are real, wall clock times; given in seconds.

#### The Stanislas Cohort

A cohort study consists of examining a given population during a period of time and of recording different data concerning this population. Data from a cohort show a high rate of complexity: they vary in time, involve a large number of individuals and parameters, show many different types, e.g. quantitative, qualitative, textual, binary, etc., and they may be noisy or incomplete.

The STANISLAS cohort is a ten-year family study whose main objective is to investigate the impact of genetic and environmental factors on variability of cardiovascular risk factors [2]. The cohort consists of 1006 presumably healthy families (4295 individuals) satisfying some criteria: French origin, two parents, at least two biological children aged of 4 or more, with members free from serious and/or chronic illnesses. The collected data are of four types: (1) Clinical data (e.g. size, weight, blood pressure); (2) Environmental data (life habits, physical activity, drug intake); (3) Biological data (glucose, cholesterol, blood count); (4) Genetic data (genetic polymorphisms).

The experts involved in the study of the STANISLAS cohort are specialists of the cardiovascular domain and they are interested in finding associations relating one or more genetic features (polymorphisms) to biological cardiovascular risk factors. The objective of the present experiment is to discover rare association

<sup>3</sup> <http://coron.loria.fr>

rules linking biological risk factors and genetic polymorphisms. As a genetic polymorphism is defined as a variation in the DNA sequence occurring in at least one percent of the population, it is easily understandable that the frequency of the different genetic variants is relatively low in the STANISLAS cohort, given that it is based on a healthy population. Therefore, this fully justifies an analysis based on rare association rules [17].

Here is an example of the extraction of a new biological hypothesis derived from the study of the STANISLAS cohort. The objective of the experiment is to characterize the genetic profile of individuals presenting “metabolic syndrome” (depending on criteria such as waist circumference, triglyceride levels, HDL cholesterol concentration, blood pressure, and fasting glucose value). A horizontal projection allowed us to retain nine individuals with metabolic syndrome. Then, a vertical projection was applied on a set of chosen attributes. Rare association rules were computed and the set of extracted rules was mined for selecting rules with the attribute *metabolic syndrome* in the left or in the right hand side. In this way, an interesting extracted rule has been discovered:  $MS \Rightarrow APOB\_71ThrIle$  (support 9 and confidence 100%). This rule can be interpreted as “an individual presenting the metabolic syndrome is heterozygous for the APOB 71Thr/Ile polymorphism”. This rule has been verified and validated using statistical tests, allowing us to conclude that the repartition of genotypes of the APOB71 polymorphism is significantly different when an individual presents metabolic syndrome or not, and suggests a new biological hypothesis: a subject possessing the rare allele for the APOB 71Thr/Ile polymorphism presents more frequently the metabolic syndrome. Other examples of rare rules can be found in [17].

### Further Experiments

We evaluated *BtB* on three more datasets. The T20I6D100K<sup>4</sup> is a sparse dataset, constructed according to the properties of market basket data that are typically sparse, weakly correlated data. The C73D10K is a census dataset from the PUMS sample file, while the MUSHROOMS<sup>5</sup> describes the characteristics of various species of mushrooms. The latter two are dense, highly correlated datasets. Table 1 shows the different steps of finding exact mRG rules. The table contains the following columns: (1) Name of the dataset; (2) Minimum support value; (3) Number of frequent itemsets. It is only indicated to show the combinatorial explosion of FIs as *min\_supp* is lowered; (4) Number of mRGs whose support exceeds 0. Since the total number of zero itemsets can be huge, we have decided to prune itemsets with support 0; (5) Number of non-singleton rare equivalence classes that are found by using non-zero mRGs; (6) Number of found exact (non-zero) mRG rules; (7) Total runtime of the *BtB* algorithm, including input/output.

During the experiments we used two limits: a space limit, which was determined by the main memory of our test machine, and a time limit that we fixed as 10,000

---

<sup>4</sup> <http://www.almaden.ibm.com/software/quest/Resources/>

<sup>5</sup> <http://kdd.ics.uci.edu/>

**Table 1.** Steps taken to find the exact mRG association rules

dataset	min_supp	# FIs	# mRGs (non-zero)	# rare eq. classes (non-zero, non-singleton)	# mRG rules (exact)	runtime of the BtB alg. (sec.)
$\mathcal{D}$	80%	5	6	3	5	0.09
T20I6D100K	10%	7	907	27	27	25.36
	0.75%	4,710	211,561	4,049	4,053	312.63
	0.5%	26,836	268,589	16,100	16,243	742.40
	<b>0.25%</b>	155,163	534,088	43,458	45,991	2,808.54
C73D10K	95%	1,007	1,622	1,570	1,622	59.10
	75%	235,271	1,939	1,794	1,939	2,183.70
	70%	572,087	2,727	2,365	2,727	4,378.02
	<b>65%</b>	1,544,691	3,675	2,953	3,675	9,923.94
MUSHROOMS	50%	163	147	139	147	3.38
	10%	600,817	2,916	2,324	2,916	74.60
	5%	4,137,547	7,963	5,430	7,963	137.86
	1%	92,894,869	37,034	16,799	37,034	321.78

seconds. The value of the barrier is printed in bold in Table 1. For instance, in the database C73D10K using *Apriori* we were unable to extract any association rules with support lower than 65% because of hitting the time limit. However, changing to *BtB* at this *min\_supp* value, we managed to extract 3,675 exact mRG rules whose supports are *below* 65%. This result shows that our method is capable to find rare rules where frequent itemset mining algorithms fail.

## 4 Conclusion

Frequent association rule mining has been studied extensively in the past. The model used in all these studies, however, has always been the same, i.e. finding all rules that satisfy user-specified *min\_supp* and *min\_conf* constraints. However, in many cases, most rules with high support are obvious and/or well-known, and it is the rules of low support that provide interesting new insights.

In this paper we presented a novel method to extract interesting rare association rules that remain *hidden* for conventional frequent itemset mining algorithms. To the best of our knowledge, this is the first method in the literature that can find strong but rare associations, i.e., local regularities in the data. These rules, called “mRG rules”, have two merits. First, they are maximally informative in the sense that they have an antecedent which is a generator itemset whereas adding the consequent to it yields a closed itemset. Second, the number of these rules is minimal, i.e. the mRG rules constitute a compact representation of all highly confident associations that can be drawn from the minimal rare itemsets.

## References

1. Liu, H., Lu, H., Feng, L., Hussain, F.: Efficient Search of Reliable Exceptions. In: Zhong, N., Zhou, L. (eds.) PAKDD 1999. LNCS (LNAI), vol. 1574, pp. 194–203. Springer, Heidelberg (1999)

2. Mansour-Chemaly, M., Haddy, N., Siest, G., Visvikis, S.: Family studies: their role in the evaluation of genetic cardiovascular risk factors. *Clin. Chem. Lab. Med.* 40(11), 1085–1096 (2002)
3. Weiss, G.: Mining with rarity: a unifying framework. *SIGKDD Explor. Newsl.* 6(1), 7–19 (2004)
4. Liu, B., Hsu, W., Ma, Y.: Mining Association Rules with Multiple Minimum Supports. In: *Proc. of the 5th ACM SIGKDD Intl. Conf. on Knowledge discovery and data mining (KDD 1999)*, pp. 337–341. ACM Press, New York (1999)
5. Yun, H., Ha, D., Hwang, B., Ryu, K.: Mining association rules on significant rare data using relative support. *Journal of Systems and Software* 67(3), 181–191 (2003)
6. Koh, Y., Rountree, N.: Finding Sporadic Rules Using Apriori-Inverse. In: Ho, T.-B., Cheung, D., Liu, H. (eds.) *PAKDD 2005. LNCS (LNAI)*, vol. 3518, pp. 97–106. Springer, Heidelberg (2005)
7. Koh, Y., Rountree, N., O’Keefe, R.: Mining Interesting Imperfectly Sporadic Rules. In: Ng, W.-K., Kitsuregawa, M., Li, J., Chang, K. (eds.) *PAKDD 2006. LNCS (LNAI)*, vol. 3918, pp. 473–482. Springer, Heidelberg (2006)
8. Szathmary, L., Napoli, A., Valtchev, P.: Towards Rare Itemset Mining. In: *Proc. of the 19th IEEE Intl. Conf. on Tools with Artificial Intelligence (ICTAI 2007)*, Patras, Greece, vol. 1, pp. 305–312 (2007)
9. Wang, K., Jiang, Y., Lakshmanan, L.V.S.: Mining unexpected rules by pushing user dynamics. In: *KDD 2003: Proceedings of the ninth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 246–255. ACM, New York (2003)
10. Mannila, H., Toivonen, H.: Levelwise Search and Borders of Theories in Knowledge Discovery. *Data Mining and Knowledge Discovery* 1(3), 241–258 (1997)
11. Kryszkiewicz, M.: Concise Representations of Association Rules. In: Hand, D.J., Adams, N.M., Bolton, R.J. (eds.) *Pattern Detection and Discovery. LNCS (LNAI)*, vol. 2447, pp. 92–109. Springer, Heidelberg (2002)
12. Bastide, Y., Taouil, R., Pasquier, N., Stumme, G., Lakhal, L.: Mining Frequent Patterns with Counting Inference. *SIGKDD Explor. Newsl.* 2(2), 66–75 (2000)
13. Boulicaut, J.F., Bykowski, A., Rigotti, C.: Free-Sets: A Condensed Representation of Boolean Data for the Approximation of Frequency Queries. *Data Mining and Knowledge Discovery* 7(1), 5–22 (2003)
14. Calders, T., Rigotti, C., Boulicaut, J.F.: A Survey on Condensed Representations for Frequent Sets. In: Boulicaut, J.F., De Raedt, L., Mannila, H. (eds.) *Constraint-Based Mining and Inductive Databases. LNCS (LNAI)*, vol. 3848, pp. 64–80. Springer, Heidelberg (2006)
15. Boros, E., Gurvich, V., Khachiyan, L., Makino, K.: On the Complexity of Generating Maximal Frequent and Minimal Infrequent Sets. In: Alt, H., Ferreira, A. (eds.) *STACS 2002. LNCS*, vol. 2285, pp. 133–141. Springer, Heidelberg (2002)
16. Kryszkiewicz, M.: Representative Association Rules. In: Wu, X., Kotagiri, R., Korb, K.B. (eds.) *PAKDD 1998. LNCS*, vol. 1394, pp. 198–209. Springer, Heidelberg (1998)
17. Szathmary, L.: Symbolic Data Mining Methods with the Coron Platform. PhD Thesis in Computer Science, Univ. Henri Poincaré – Nancy 1, France (November 2006)

# Background Knowledge Integration in Clustering Using Purity Indexes

Germain Forestier\*, Cédric Wemmert, and Pierre Gançarski

Image Sciences, Computer Sciences and Remote Sensing Laboratory  
University of Strasbourg, France

{forestier,wemmert,gancarski}@unistra.fr  
<https://lsiit-cnrs.unistra.fr/>

**Abstract.** In recent years, the use of background knowledge to improve the data mining process has been intensively studied. Indeed, background knowledge along with knowledge directly or indirectly provided by the user are often available. However, it is often difficult to formalize this kind of knowledge, as it is often dependent of the domain. In this article, we studied the integration of knowledge as labeled objects in clustering algorithms. Several criteria allowing the evaluation of the purity of a clustering are presented and their behaviours are compared using artificial datasets. Advantages and drawbacks of each criterion are analyzed in order to help the user to make a choice among them.

**Keywords:** Clustering, background knowledge, semi-supervised algorithm, purity indexes.

## 1 Introduction

Knowledge integration to guide the clustering process is a major issue in data mining and an active research area. Indeed, fully unsupervised approaches raise some problems when dealing with more and more complex data. Moreover, background knowledge on the studied data are often available. Thus, it is important to work on proposing new approaches (semi-supervised methods) able to deal with such knowledge, to produce better results and to enhance the performance of the algorithms (speed-up, quality of the solutions, etc.).

The background knowledge can be represented in many different ways as they are strongly dependent on the studied domain. Even the number of clusters to find can be considered as knowledge on the data. Many works [24,5] addressed the problem of using background knowledge, represented as constraints between two objects of the dataset. These constraints give the information that the two objects have to be in the same cluster (*must-link*) or, on the contrary, that they should not be in the same cluster (*cannot-link*). Labeled samples can also

---

\* Please note that the LNCS Editorial assumes that all authors have used the western naming convention, with given names preceding surnames. This determines the structure of the names in the running heads and the author index.

be considered as another kind of knowledge. In a similar way than supervised classification methods that learn a classification function from a learning set composed of labeled objects, this information can be used during the clustering process to guide the algorithm towards a solution respecting this knowledge. It is not necessary to have many labeled samples as in the supervised case, and they do not have to belong to each class of the problem.

In this context, the concept of *purity* of the clusters is very important. The purity evaluates the quality of the clusters according to the labeled samples available. A cluster is considered pure if it contains labeled objects from one and only one class. Inversely, a cluster is considered as impure if it contains labeled objects from many different classes.

The purpose of this article is to present and compare many different ways to evaluate the purity of the clusters. In the section 2, we give a state of the art about knowledge integration in data mining to introduce the context of this study. Then, in section 3, purity indexes are formalized and compared. Finally, we draw conclusions and give some directions of future work.

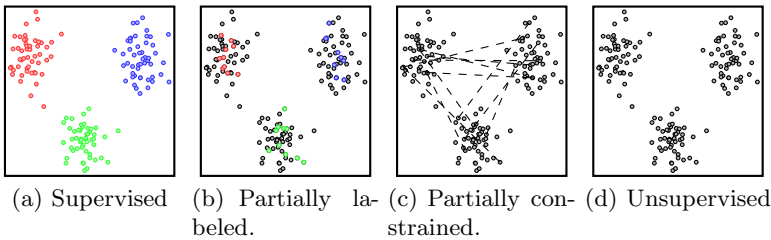


Fig. 1. Example of different kinds of background knowledge

## 2 Clustering with Background Knowledge

Many approaches have been investigated to use background knowledge to guide the clustering process.

In constrained clustering, knowledge is expressed as *must-link* and *cannot-link* constraints and is used to guide the clustering process. A *must-link* constraint gives the information that two data objects should be in the same cluster, and *cannot-link* means the opposite. This kind of knowledge is sometimes easier to obtain than a classical subset of labeled samples. Wagstaff et al. [24] presented a constrained version of the KMEANS algorithm which uses such constraints to bias the assignment of the objects to the clusters. At each step, the algorithm tries to agree with the constraints given by the user. These constraints can also be used to learn a distance function biased by the knowledge about the links between the data objects [5]. The distance between two data objects is reduced for a must-link and increased for a cannot-link. Huang et al. [14] presented an active learning framework for semi-supervised document clustering with language modeling. The approach uses a gain-directed document pair selection method to



select cleverly the constraints. In order to minimize the amount of constraints required, Griga et al. [13] defined an active mechanism for the selection of candidate constraints. The active fuzzy constrained clustering method is presented and evaluated on a ground truth image database to illustrate that the clustering can be significantly improved with few constraints. Recent works on constrained clustering are focused on evaluating the utility (i.e. the potential interest) of a set of constraints [7,25].

Kumar and Kummamuru [16] introduced another kind of knowledge through a clustering algorithm that uses supervision in terms of relative comparisons, e.g.  $x$  is closer to  $y$  than to  $z$ . Experimental studies on high-dimensional textual data sets demonstrated that the proposed algorithm achieved higher accuracy and is more robust than similar algorithms using pairwise constraints (*must-link* and *cannot-link*) for supervision. Klein et al. [15] allowed instance-level constraints (i.e. *must-link*, *cannot-link*) to have space level inductive implications in order to improve the use of the constraints. This approach improved the results of the previously studied constrained KMEANS algorithms and generally requires less constraints to obtain the same accuracies. Basu et al. [3] presented a pairwise constrained clustering framework as well as a new method for actively selecting informative pairwise constraints, to get improved clustering performance. Experimental and theoretical results confirm that this active querying of pairwise constraints significantly improves the accuracy of clustering, when given a relatively small amount of supervision.

Another way to integrate background knowledge is to use a small set of labeled samples. Basu et al. [2] used a set of samples to *seed* (i.e. to initialize) the clusters of the KMEANS algorithm. Two algorithms, SEEDED-KMEANS and CONSTRAINED-KMEANS, are presented. In the first one, the labeled samples are used to initialize the clusters and the clusters are updated during the clustering process such as in the KMEANS algorithm. In the second one, the labeled samples used during the initialization stay in their assigned cluster, and only the unlabeled samples can change cluster during the cluster allocation step of KMEANS. The choice between these two approaches must be done according to the knowledge about noise in the dataset.

To tackle the problem of incorporating partial background knowledge into clustering, when the labeled samples have moderate overlapping features with the unlabeled data, Gao et al. [12] formulated a new approach as a constrained optimization problem. The authors introduced two learning algorithms to solve the problem, based on hard and fuzzy clustering methods. An empirical study shows that the proposed algorithms improve the quality of clustering results despite a limited number of labeled samples. Basu et al. [4] also proposed a probabilistic model for semisupervised clustering, based on HIDDEN MARKOV RANDOM FIELDS (HMRF), that provides a principled framework for incorporating supervision into prototype-based clustering. Experimental results on several text data sets demonstrate the advantages of this framework.

Another approach, called supervised clustering [10], uses the class information about the objects as an additional feature, to build clusters with a high

class-based purity. The goal of supervised clustering is to identify class-uniform clusters having high probability densities. Supervised clustering is used to create summaries of datasets and for enhancing existing classification algorithms.

Different kinds of background knowledge are introduced by Pedrycz et al. [19], namely partial supervision, proximity-based guidance and uncertainty driven knowledge hints. The authors discuss different ways of exploiting and effectively incorporating these background knowledge (known as *knowledge hints*) in the fuzzy c-means algorithm. In [6], Bouchachia and Pedrycz presented an extension of the fuzzy collaborative clustering which consists in taking into account background knowledge through labeled objects. One of the advantages of the method is to take into account the classes split in several clusters. During the collaboration step, the method identifies if a class corresponds to various clusters and add or remove clusters according to this information. More recently, Pedrycz [20] presented some concepts and algorithms on collaborative and knowledge-based fuzzy clustering. The FUZZY C-MEANS algorithm (FCM) was used as an operational model to explain the approach. Interesting linkages between information granularity, privacy and security of data in collaborative clustering were also discussed. The problem of data privacy when clustering multiple datasets was also recently discussed in [11]. An application of fuzzy clustering with partial knowledge to organize and classify digital images is also proposed in [17]. The author present an operational framework of fuzzy clustering using the FUZZY C-MEANS algorithm with an augmented objective function using background knowledge. Experiments are carried out on collections of images composed of 2000 images.

### 3 Clustering Evaluation

The evaluation of the purity or the quality of the partition produced by a clustering consists in determining if the repartition of the objects in the different clusters is coherent with the available knowledge on the data. We consider here the knowledge as a set of labeled objects. Let us define some notations to formalize the purity indexes:

- Let  $N$  be the number of labeled samples
- Let  $\mathbb{C} = \{c_1, c_2, \dots, c_K\}$  be the clusters found by the clustering algorithm
- Let  $\mathbb{W} = \{w_1, w_2, \dots, w_C\}$  be the classes of the labeled objects
- Let  $c_k$  be the objects composing cluster  $k$  and  $w_k$  the objects composing class  $k$
- Let  $|c_k|$  be the cardinality of cluster  $k$
- Let  $n_j^i = |c_i \cap w_j|$  be the number of objects of cluster  $i$  being in class  $j$ .

#### 3.1 Purity Evaluation

The easiest way to compute the purity of a clustering is to find the majority class in each cluster and to count the number of labeled objects of this class in each cluster [18]. Then, the purity can be defined as:

$$\mathbf{H}_{\text{simple}}(\mathbb{C}, \mathbb{W}) = \frac{1}{N} \sum_i^K \arg \max_j (n_j^i) \quad (1)$$

This evaluation of the purity consists in estimating the percentage of labeled objects of the majority class in each cluster of the clustering. It takes its value in  $[0; 1]$ , 1 indicating that all clusters are pure, i.e. they contain only labeled objects of one class.

Another way to estimate the clusters purity is proposed by Solomonoff et al. [23]. The authors define the purity as the probability that, given a cluster  $i$  and two randomly chosen labeled objects of this cluster, they both are of the same class  $j$ . The probability that the class of the first object is  $j$  is  $n_j^i/|c_i|$ . The probability that the class of the second object is also  $j$  is  $(n_j^i/|c_i|)^2$ . Finally, the purity of a cluster  $i$  can be defined as:

$$\pi_{\text{prob}}(c_i) = \sum_j^C \left( \frac{n_j^i}{|c_i|} \right)^2 \quad (2)$$

which can be derived to a clustering by:

$$\mathbf{H}_{\text{prob}}(\mathbb{C}, \mathbb{W}) = \frac{1}{N} \sum_i^K |c_i| \pi_{\text{prob}}(c_i) \quad (3)$$

The advantage of this measure, compared to the simple purity evaluation (1), is that it takes into account the distribution of all the classes in the cluster, and not only the majority class. Thus, it promotes clusters composed of labeled samples from a limited number of classes. Its value is in  $[0; 1]$ , 1 indicating that all the clusters are pure.

However, these two purity indexes present a major drawback. They over evaluate the quality of a clustering having a large number of clusters. Indeed, the purity is maximal when having one cluster per objects (which is generally not considered as a *good* solution). De facto, if these measures are used in an algorithm allowing the number of clusters to change, it will tend to converge to a solution having to many clusters. Many propositions have been given to cope with this problem. In [1], Ajmera et al. have proposed to calculate the clusters purity according to their composition in terms of classes, but also the purity of the classes in terms of clusters (for each class, its distribution among all the clusters is observed). Then, the two values are merged to become the purity evaluation of the clustering. This enables to penalize solutions proposing too many clusters. The classes purity is computed in the same way as the clusters purity:

$$\pi_{\text{prob}}^{\sim}(w_i) = \sum_j^K \left( \frac{n_j^i}{|w_i|} \right)^2 \quad (4)$$

which gives the following definition for all the clusters of a clustering:

$$\mathbf{H}_{\text{prob}}^{\sim}(\mathbb{C}, \mathbb{W}) = \frac{1}{N} \sum_i^K |w_i| \pi_{\text{prob}}^{\sim}(w_i) \quad (5)$$

The clusters purity and the classes purity are then combined as follows:

$$\mathbf{II}_{\text{overall}}(\mathbb{C}, \mathbb{W}) = \sqrt{\mathbf{II}_{\text{prob}}(\mathbb{C}, \mathbb{W}) \times \mathbf{II}_{\text{prob}}^{\sim}(\mathbb{C}, \mathbb{W})} \quad (6)$$

Another approach consists in also considering a quality measure of the clustering. Demiriz et al. [9] used an optimization algorithm with a purity index called GINI which is similar to the criterion given in 2. To avoid this case, where the algorithm generates solutions with too many clusters, the objective function to optimize is an arithmetic mean between the clusters purity and quality. The quality of the clusters is evaluated according to Davies-Bouldin index [8], which promotes well separated compact clusters. The combination of these two criteria enables to avoid extreme solutions (e.g. one cluster for each object).

Finally, Eick et al. [10] proposed to use a penalty criterion, to penalize solutions having too many clusters. The penalty is calculated as follows:

$$\text{penalty}(K) = \begin{cases} \sqrt{\frac{K-C}{N}} & \text{if } K \geq C \\ 0 & \text{otherwise} \end{cases} \quad (7)$$

with  $K$  the number of clusters,  $C$  the number of classes and  $N$  the number of objects. It can be used with any purity index, as the simple criteria (1):

$$\mathbf{II}_{\text{penalty}}(\mathbb{C}, \mathbb{W}) = \mathbf{II}_{\text{simple}}(\mathbb{C}, \mathbb{W}) - \beta \text{penalty}(K) \quad (8)$$

Another solution is to evaluate the Normalized Mutual Information (NMI) index between available knowledge and the clustering:

$$\mathbf{II}_{\text{nmi}}(\mathbb{C}, \mathbb{W}) = \frac{I(\mathbb{C}, \mathbb{W})}{[H(\mathbb{C}) + H(\mathbb{W})]/2} \quad (9)$$

$I$  is the mutual information:

$$I(\mathbb{C}, \mathbb{W}) = \sum_i \sum_j \frac{n_j^i}{N} \log \frac{n_j^i/N}{|c_i|/N \times |w_j|/N} \quad (10)$$

$$= \sum_i \sum_j \frac{n_j^i}{N} \log \frac{N \times n_j^i}{|c_i| \times |w_j|} \quad (11)$$

$H$  is the entropy:

$$H(\mathbb{W}) = - \sum_k \frac{|w_k|}{N} \log \frac{|w_k|}{N} \quad (12)$$

The mutual information  $I$  (10) evaluates the quantity of information provided by the clustering on the classes. The denominator in (9) enables to normalize the criterion which value is in  $[0; 1]$ , 1 indicating pure clusters. This index is maximal when the number of clusters is equal to the number of classes. Thus, it does not have the drawback of the previous criteria presented above.

### 3.2 Partitions Comparison

Another commonly used criterion to compare partitions is the *rand* index [21]. It consists in comparing pairs of objects and to check if they are classified identically in two partitions. In our case, we verify if each pair of objects identically labeled according to the background knowledge are in the same cluster. A pair of objects is a *true positive* (TP) if the two objects have the same label and are in the same cluster. It is a *true negative* (TN) if they have different labels and are in different clusters. A *false positive* (FP) corresponds to a pair of objects having different labels but in the same cluster, whereas a *false negative* (FN) corresponds to a pair of objects having the same label but being in two different clusters. The *rand* index can then be defined as:

$$\mathbf{II}_{\text{rand}}(\mathbb{C}, \mathbb{W}) = \frac{TP + TN}{TP + FP + FN + TN} \quad (13)$$

$(TP + FP + FN + TN)$  representing all pairs of objects and  $(TP + TN)$  all pairs of objects correctly classified. One of the drawbacks of this index, is that the same weight is given to false positives and false negatives (another is, e.g., it's poor behavior in case of the unbalanced classes).

Regarding the *F-Measure* [22], it enables to affect weights to these values, according to the precision (P) and the recall (R):

$$P = \frac{TP}{TP + FP} \quad R = \frac{TP}{TP + FN}$$

$$\mathbf{II}_{\text{fmeasure}}(\mathbb{C}, \mathbb{W}) = \frac{(\beta^2 + 1)P \times R}{\beta^2 P + R} \quad (14)$$

The  $\beta$  parameter can be used to more penalize the false negatives as the false positives, giving it a value over one ( $\beta > 1$ ). If  $\beta = 1$ , the precision and recall have the same importance.

The advantage of these two criteria ( $\mathbf{II}_{\text{rand}}$  and  $\mathbf{II}_{\text{fmeasure}}$ ) is that they implicitly integrate the number of clusters, putting the solutions proposing too many clusters at a disadvantage. Indeed, the more the number of clusters is increased, the more the pairs of objects differ from the available knowledge.

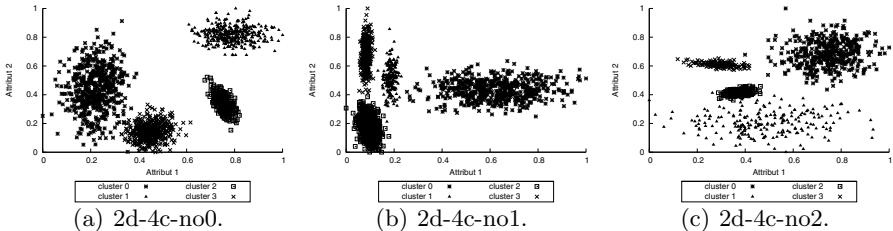
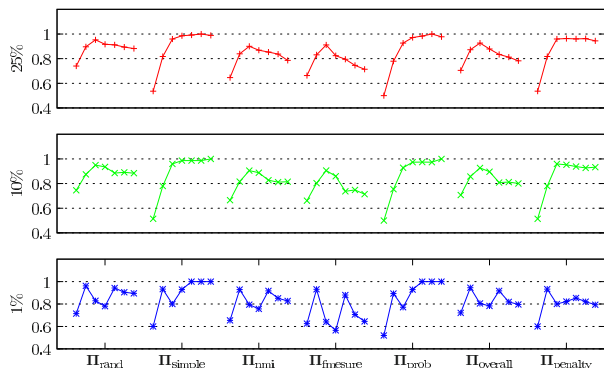
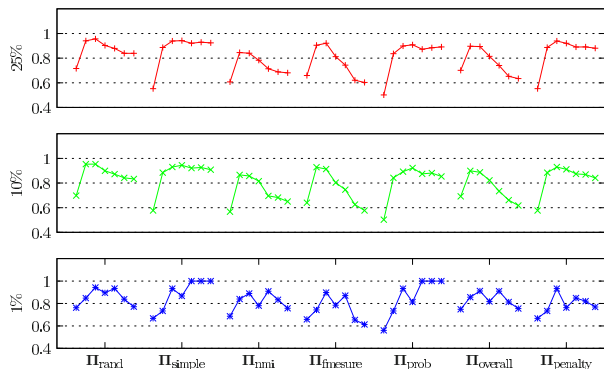


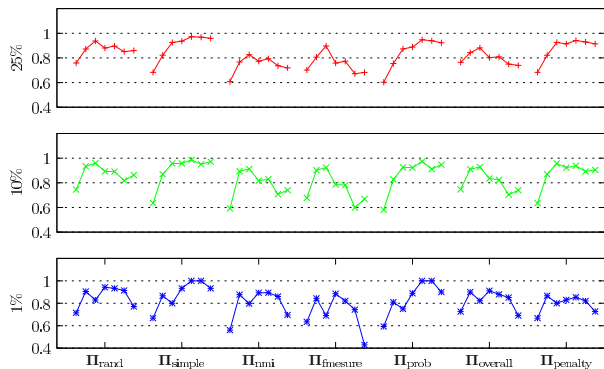
Fig. 2. The three datasets used for the evaluation



(a) Criteria evolution according to the number of clusters for the dataset Fig. 2(a)



(b) Criteria evolution according to the number of clusters for the dataset Fig. 2(b).



(c) Criteria evolution according to the number of clusters for the dataset Fig. 2(c).

**Fig. 3.** Criteria evolution

### 3.3 Evaluation of the Quality Criteria

In this section, the criteria presented before are going to be evaluated on different datasets. Figure 2 shows three artificial datasets, each representing four clusters in a two dimension space. The KMEANS algorithm was used on these data, with a number of clusters varying from 2 to 8. For each clustering, the different measures presented in the previous sections were calculated. Three configurations have been evaluated, the first with 1% of labeled objects, the second with 10% of labeled objects and the last one with 25% of labeled objects in the dataset. Each experiment has been ran 100 times with random initialization, and the results were averaged. Figures 3 (a), (b) and (c ) give the results respectively for the dataset presented in figure 2(a), 2(b) and 2(c).

One can observe that when only few labeled objects are available (1%), almost all the criteria have an unpredictable behaviour. Indeed, it is not guaranteed that the labeled set contains examples of all classes. That is why these criteria can hardly be used when only limited knowledge is available. When the number of labeled objects increases (10%), the probability to have samples of each class in the labeled set also increases. Therefore, the evolutions of the criteria are more typical. One can observe the already mentioned problem that some purity measures overevaluate the quality of the clustering when the number of clusters increase. Indeed, the simple purity index ( $\mathbf{II}_{\text{simple}}$ ) and the cluster purity index ( $\mathbf{II}_{\text{prob}}$ ) increase as the number of clusters increase. The other criteria ( $\mathbf{II}_{\text{rand}}$ ,  $\mathbf{II}_{\text{nmi}}$ ,  $\mathbf{II}_{\text{fmeasure}}$ ,  $\mathbf{II}_{\text{overall}}$ ,  $\mathbf{II}_{\text{penalty}}$ ) tend to decrease as the number of clusters increase. The most characteristic are  $\mathbf{II}_{\text{fmeasure}}$ ,  $\mathbf{II}_{\text{overall}}$  and the  $\mathbf{II}_{\text{nmi}}$ . The criteria  $\mathbf{II}_{\text{rand}}$  and  $\mathbf{II}_{\text{penalty}}$  decrease less significantly. It is interesting to notice that there is no noticeable difference between the results obtained with 10% or 25% of labeled objects.

## 4 Conclusion

Knowledge integration in clustering algorithms is a really important issue. As more and more knowledge are available on the data manipulated, it is necessary to propose new approaches that enables to deal with this huge amount of information.

In this article, we have presented how to take advantage of labeled objects to evaluate the purity of a clustering. Many criteria were exhibited, formalized and compared. One observation is that purity evaluation without taking into account the number of clusters tends to overevaluate the quality of the results. To cope with this problem, it is possible to penalize results with a huge number of clusters. Another type of criteria only compare how pairs of objects were classified, as the *F-Measure* which has given particularly good results in our experiments.

In the future, we aim to evaluate more criteria and to compare other types of domain knowledge, as for example constraints on the objects of the dataset. Moreover, it would be necessary to study the behaviour of these criteria when labeled objects of the same class belong to different clusters.

## References

1. Ajmera, J., Boulard, H., Lapidot, I., McCowan, I.: Unknown-multiple speaker clustering using hmm. In: International Conference on Spoken Language Processing, September 2002, pp. 573–576 (2002)
2. Basu, S., Banerjee, A., Mooney, R.J.: Semi-supervised clustering by seeding. In: International Conference on Machine Learning, pp. 19–26 (2002)
3. Basu, S., Banerjee, A., Mooney, R.J.: Active semi-supervision for pairwise constrained clustering. In: SIAM International Conference on Data Mining, pp. 333–344 (2004)
4. Basu, S., Bilenko, M., Mooney, R.J.: A probabilistic framework for semi-supervised clustering. In: International Conference on Knowledge Discovery and Data Mining, pp. 59–68 (2004)
5. Bilenko, M., Basu, S., Mooney, R.J.: Integrating constraints and metric learning in semi-supervised clustering. In: International Conference on Machine Learning, pp. 81–88 (2004)
6. Bouchachia, A., Pedrycz, W.: Data clustering with partial supervision. *Data Min. Knowl. Discov.* 12(1), 47–78 (2006)
7. Davidson, I., Wagstaff, K.L., Basu, S.: Measuring constraint-set utility for partitioning clustering algorithms. In: European Conference on Principles and Practice of Knowledge Discovery in Databases, pp. 115–126 (2006)
8. Davies, D., Bouldin, D.: A cluster separation measure. *IEEE Transactions on Pattern Analysis and Machine Intelligence PAMI-1*(2), 224–227 (1979)
9. Demiriz, A., Bennett, K., Embrechts, M.: Semi-supervised clustering using genetic algorithms. In: Intelligent Engineering Systems Through Artificial Neural Networks, pp. 809–814 (1999)
10. Eick, C.F., Zeidat, N., Zhao, Z.: Supervised clustering - algorithms and benefits. In: International Conference on Tools with Artificial Intelligence, pp. 774–776 (2004)
11. Fung, B.C., Wang, K., Wang, L., Hung, P.C.: Privacy-preserving data publishing for cluster analysis. *Data & Knowledge Engineering* 68(6), 552–575 (2009)
12. Gao, J., Tan, P., Cheng, H.: Semi-supervised clustering with partial background information. In: SIAM International Conference on Data Mining, pp. 489–493 (2006)
13. Grira, N., Crucianu, M., Boujemaa, N.: Active semi-supervised fuzzy clustering. *Pattern Recognition* 41(5), 1851–1861 (2008)
14. Huang, R., Lam, W.: An active learning framework for semi-supervised document clustering with language modeling. *Data & Knowledge Engineering* 68(1), 49–67 (2009)
15. Klein, D., Kamvar, S., Manning, C.: From instance-level constraints to space-level constraints: Making the most of prior knowledge in data clustering. In: The Nineteenth International Conference on Machine Learning, pp. 307–314 (2002)
16. Kumar, N., Kummamuru, K.: Semisupervised clustering with metric learning using relative comparisons. *IEEE Transactions on Knowledge and Data Engineering* 20(4), 496–503 (2008)
17. Loia, V., Pedrycz, W., Senatore, S.: Semantic web content analysis: A study in proximity-based collaborative clustering. *IEEE Transactions on Fuzzy Systems* 15(6), 1294–1312 (2007)
18. Manning, C.D., Raghavan, P., Schütze, H.: *Introduction to Information Retrieval*. Cambridge University Press, Cambridge (2008)
19. Pedrycz, W.: Fuzzy clustering with a knowledge-based guidance. *Pattern Recognition Letters* 25(4), 469–480 (2004)



20. Pedrycz, W.: Collaborative and knowledge-based fuzzy clustering. *International Journal of Innovative, Computing, Information and Control* 1(3), 1–12 (2007)
21. Rand, W.M.: Objective criteria for the evaluation of clustering methods. *Journal of the American Statistical Association* 66, 622–626 (1971)
22. van Rijsbergen, C.J.: *Information Retrieval*. Butterworths, London (1979)
23. Solomonoff, A., Mielke, A., Schmidt, M., Gish, H.: Clustering speakers by their voices. In: *Proceedings of the 1998 IEEE International Conference on Acoustics, Speech and Signal Processing*, May 1998, vol. 2, pp. 757–760 (1998)
24. Wagstaff, K., Cardie, C., Rogers, S., Schroedl, S.: Constrained k-means clustering with background knowledge. In: *International Conference on Machine Learning*, pp. 557–584 (2001)
25. Wagstaff, K.L.: Value, cost, and sharing: Open issues in constrained clustering. In: Dzeroski, S., Struyf, J. (eds.) *KDID 2006. LNCS*, vol. 4747, pp. 1–10. Springer, Heidelberg (2007)

# A Comparison of Merging Operators in Possibilistic Logic

Guilin Qi<sup>1\*</sup>, Weiru Liu<sup>2</sup>, and David Bell<sup>2</sup>

<sup>1</sup> School of Computer Science and Engineering, Southeast University, Nanjing 210096

<sup>2</sup> School of Electronics, Electrical Engineering and Computer Science  
Queen's University Belfast, Belfast, BT7 1NN, UK

**Abstract.** In this paper, we compare some important merging operators in possibilistic logic. We first introduce semantic merging operators and adaptive merging operators in possibilistic logic. We then propose an approach to evaluating the *discriminating power* of these merging operators. After that, we analyze the computational complexity of these possibilistic merging operators. Finally, we consider the compatibility of possibilistic merging operators with propositional merging operators.

## 1 Introduction

Fusion of information coming from different sources is crucial to build intelligent systems. In classical logic, this problem is often called belief merging, which defines the beliefs (resp. goals) of a group of agents from their individual beliefs (resp. goals). It is well-known that priorities or orderings (either implicit or explicit) play an important role in belief merging [19,20,23]. The handling of priorities has been shown to be completely in agreement with possibilistic logic [14]. Possibilistic logic is a weighted logic which attaches to each first-order logic formula a weight belonging to a totally ordered scale, such as  $(0, 1]$ . An ordering between two formulas is then obtained by comparing the weights attached to them. Possibilistic logic is also known to be a good logical framework for reasoning under inconsistency and uncertainty when only partial information is available.

Many approaches have been proposed for merging uncertain information in possibilistic logic. In the framework of possibilistic logic, each source of uncertain information is represented as a *possibilistic knowledge base*, which is a set of weighted formulas. A possibilistic knowledge base has a unique possibility distribution associated with it. In [3,4], some semantic merging operators were proposed to aggregate *possibility distributions* of original possibilistic knowledge bases, the result is a new possibility distribution. Then the syntactical counterpart of a semantic merging operator is applied to the possibilistic bases, and the result of merging is a possibilistic knowledge base whose possibility distribution is the one obtained by the semantic merging operator. There are two important

---

\* Guilin Qi is partially supported by Excellent Youth Scholars Program of Southeast University under grant 4009001011.

classes of merging operators, one class contains t-norm operators (for example, the minimum operator) and the other consists of t-conorm operators (for example, the maximum operator). Several adaptive merging rules have been proposed to integrate both the conjunctive and disjunctive operators (e.g., [8,9,10]). In practice, an important problem is the choice of an appropriate merging approach. To facilitate the choice among different merging approaches, we need some criteria to evaluate the merits and drawbacks of each approach.

In this paper, we propose three criteria to evaluate merging operators in possibilistic logic. The first criterion is to evaluate the discriminating power of a merging operator, which is measured by the *nonspecificity function*. After that, we analyze the computational complexity of existing possibilistic merging operators. Finally, we consider the compatibility of possibilistic merging operators with propositional merging operators.

## 2 Preliminaries

In this paper, we consider a propositional language  $\mathcal{L}_{PS}$  from a finite set  $PS$  of propositional symbols. The classical consequence relation is denoted as  $\vdash$ . An interpretation is a total function from  $PS$  to  $\{true, false\}$ .  $\Omega$  is the set of all possible interpretations. An interpretation  $\omega$  is a model of a formula  $\phi$ , denoted  $\omega \models \phi$ , iff  $w(\phi) = true$ . A *classical knowledge base*  $B$  is a finite set of propositional formulas.  $B$  is consistent iff there exists an interpretation  $\omega$  such that  $\omega(\phi) = true$  for all  $\phi \in B$ .

**Possibilistic Logic:** Possibilistic logic [14] is a weighted logic where each classical logic formula is associated with a number in  $(0, 1]$ . A possibilistic knowledge base (or PKB for short) is a set of possibilistic formulas of the form  $B = \{(\phi_i, a_i) : i = 1, \dots, n\}$ . *Possibilistic formula*  $(\phi_i, a_i)$  means that the necessity degree of  $\phi_i$  is at least equal to  $a_i$ . A classical knowledge base  $B = \{\phi_i : i = 1, \dots, n\}$  corresponds to a PKB  $B' = \{(\phi_i, 1) : i = 1, \dots, n\}$ . A *possibilistic knowledge profile*  $\mathcal{E}$  is a multi-set of PKBs. In this paper, we consider only PKBs where every formula  $\phi$  is a classical propositional formula. The classical base associated with  $B$  is denoted as  $B^*$ , namely  $B^* = \{\phi_i | (\phi_i, a_i) \in B\}$ . A PKB is consistent iff its classical base is consistent.

The semantics of possibilistic logic is based on the notion of a *possibility distribution*  $\pi: \Omega \rightarrow [0, 1]$ .  $\pi(\omega)$  represents the possibility degree of interpretation  $\omega$  with available beliefs. A possibility distribution  $\pi$  is *normal* iff there exists  $\omega \in \Omega$  such that  $\pi(\omega) = 1$ . Given a PKB  $B$ , a unique *possibility distribution*, denoted by  $\pi_B$ , can be obtained by the principle of minimum specificity [14]. For all  $\omega \in \Omega$ ,

$$\pi_B(\omega) = \begin{cases} 1 & \text{if } \forall (\phi_i, a_i) \in B, \omega \models \phi_i, \\ 1 - \max\{a_i | \omega \not\models \phi_i, (\phi_i, a_i) \in B\} & \text{otherwise.} \end{cases} \quad (1)$$

It has been shown that a PKB  $B$  is consistent iff  $\pi_B$  is normal.

Given a PKB  $B$  and  $a \in (0, 1]$ , the  $a$ -cut of  $B$  is  $B_{\geq a} = \{\phi \in B^* | (\phi, b) \in B \text{ and } b \geq a\}$ . The *inconsistency degree* of  $B$ , denoted  $Inc(B)$ , is defined as  $Inc(B) = \max\{a_i : B_{\geq a_i} \text{ is inconsistent}\}$ .

**Definition 1.** Let  $B$  be a PKB. A formula  $\phi$  is said to be a possibilistic consequence of  $B$  to degree  $a$ , denoted by  $B \vdash_{\pi}(\phi, a)$ , iff the following conditions hold: (1)  $B_{\geq a}$  is consistent; (2)  $B_{>a} \vdash \phi$ ; (3)  $\forall b > a, B_{>b} \not\vdash \phi$ .

### 3 Merging Approaches in Possibilistic Logic

According to [1], there are two different categories of approaches for merging PKBs. The first category of approaches resolves inconsistency after merging and result in a unique consistent base (e.g. [3,4,6,25,26]). By contrast, the second category of approaches tolerates inconsistency and cope with them [?,5,?]. Because of page limit, we consider only some important merging approaches belonging to the first category.

Given  $n$  PKBs  $B_1, \dots, B_n$ , a semantic combination operator  $\oplus$  maps possibility distributions  $\pi_1, \dots, \pi_n$  into a new possibility distribution. The semantic combination can be performed easily when  $\oplus$  is associative. That is, we have  $\pi_{\oplus}(\omega) = (\dots((\pi_1(\omega) \oplus \pi_2(\omega)) \oplus \pi_3(\omega)) \oplus \dots) \oplus \pi_n(\omega)$ . When the operator is not associative, it needs to be generalized as an unary operator defined on a vector  $(\pi_1, \dots, \pi_n)$  of possibility distributions such that:

1.  $\oplus(1, \dots, 1) = 1$ , and
2. if  $\forall i = 1, \dots, n, \pi_i(\omega) \geq \pi_i(\omega')$  then  $\oplus(\pi_1(\omega), \dots, \pi_n(\omega)) \geq \oplus(\pi_1(\omega'), \dots, \pi_n(\omega'))$ .

Two classes of aggregation operator which are commonly used are t-norm (denoted as  $tn$ ) and t-conorm (denoted as  $ct$ ). Examples of t-norms are the minimum operator and the product operator and examples of t-conorms are the maximum operator and the “probabilistic sum” operator defined by  $a \oplus b = a + b - ab$ .

The merged possibility distribution of a t-norm operator may be not normal. In that case, we may think of renormalizing  $\pi_{tn}$ . Let  $\pi$  be a possibility distribution which is not normal,  $\pi_N$  be the possibility distribution renormalized from  $\pi$ . Then  $\pi_N$  should satisfy the following conditions:

1.  $\exists \omega, \pi_N(\omega) = 1$ ,
2. if  $\pi$  is normal then  $\pi_N = \pi$ ,
3.  $\forall \omega, \omega', \pi(\omega) < \pi(\omega')$  if and only if  $\pi_N(\omega) < \pi_N(\omega')$ .

For example, let  $h(\pi_{tn}) = \max_{\omega \in \Omega} \{\pi_{tn}(\omega)\}$ , the following equation provides a normalization rule:

$$\pi_{N,tn}(\omega) = \begin{cases} 1 & \text{if } \pi_{tn}(\omega) = h(\pi_{tn}), \\ \pi_{tn}(\omega) & \text{otherwise.} \end{cases} \tag{2}$$

The normalization rule defined by Equation (2) resolves inconsistency because the inconsistency degree of any PKB associated with  $\pi_{N,tn}$  is zero. Other normalization rules can be found in [3].

The syntactic generalization for a semantic operator can be carried out as follows.

**Proposition 1.** [4] Let  $\mathcal{E} = \{\mathcal{B}_1, \dots, \mathcal{B}_n\}$  be a set of  $n$  PKBs and  $(\pi_1, \dots, \pi_n)$  be their associated possibility distributions. Let  $\pi_{\mathcal{B}_{\oplus}}$  be the result of combining  $(\pi_1, \dots, \pi_n)$  with  $\oplus$ . The PKB associated with  $\pi_{\mathcal{B}_{\oplus}}$  is:

$$\mathcal{B}_{\oplus} = \{(D_j, 1 - \oplus(x_1, \dots, x_n)) : j = 1, \dots, n\}, \quad (3)$$

where  $D_j$  are disjunctions of size  $j$  between formulas taken from different  $\mathcal{B}_i$ 's ( $i = 1, \dots, n$ ) and  $x_i$  is equal to  $1 - a_i$  if  $\phi_i$  belongs to  $D_j$  and 1 if it does not.

By Equation 3, the PKBs, which are the syntactical counterparts of semantic merging using a t-norm  $tn$  and a t-conorm  $ct$  are the following knowledge bases respectively [3]:

$$\begin{aligned} \mathcal{B}_{tn} = B_1 \cup B_2 \cup \{(\phi_i \vee \psi_j, ct(a_i, b_j)) | (\phi_i, a_i) \in B_1 \\ \text{and } (\psi_j, b_j) \in B_2\}, \end{aligned} \quad (4)$$

$$\mathcal{B}_{ct} = \{(\phi_i \vee \psi_j, tn(a_i, b_j)) | (\phi_i, a_i) \in B_1, (\psi_j, b_j) \in B_2\}. \quad (5)$$

By Equation (4), the PKB  $\mathcal{B}_{tn}$  may be inconsistent. Let  $\pi_{N,tn}$  be the possibility distribution obtained by Equation (2), then the PKB associated with it has the following form:

$$\mathcal{B}_{N,tn} = \{(\phi_i, a_i) : (\phi_i, a_i) \in \mathcal{B}_{tn} \text{ and } a_i > Inc(\mathcal{B}_{tn})\}. \quad (6)$$

$\mathcal{B}_{N,tn}$  restores consistency of  $\mathcal{B}_{tn}$  by dropping formulas whose weights are less than or equal to the inconsistency degree of  $\mathcal{B}_{tn}$ . We call the merging operator obtained by Equation 6 a normalized conjunctive merging operator. It is clear that  $\mathcal{B}_{N,tn}$  may drop too much information from  $\mathcal{B}_{tn}$  if  $Inc(\mathcal{B}_{tn})$  is large, for example, 0.8. This is because possibilistic inference suffers from the *drowning problem* [2], which says that formulas whose weights are less than or equal to the inconsistency degree of a PKB are not useful for the inference.

*Example 1.* Let  $B_1 = \{(p, 0.9), (q, 0.7)\}$  and  $B_2 = \{(\neg p, 0.8), (r, 0.6), (p \vee q, 0.5)\}$ . Suppose the operator is the maximum, then by Equation 5, we have  $\mathcal{B}_{max} = \{(p \vee r, 0.6), (p \vee q, 0.5), (\neg p \vee q, 0.7), (q \vee r, 0.6)\}$ . It is clear that the maximum based merging operator is very *cautious*, that is, all the formulas are weakened as disjunctions. By contrast, if we choose the minimum, then by Equation 4, we have  $\mathcal{B}_{min} = \{(p, 0.9), (\neg p, 0.8), (q, 0.7), (r, 0.6), (p \vee q, 0.5)\}$ .  $\mathcal{B}_{min}$  is inconsistent. Suppose we apply the normalization rule (Equation (2)) to the possibility distribution associated with  $\mathcal{B}_{min}$ , then by Equation 6 the PKB associated with the normalized possibility distribution is  $\mathcal{B}_{N,min} = \{(p, 0.9)\}$ .  $(q, 0.7)$  and  $(r, 0.6)$  are not involved in conflict between  $B_1$  and  $B_2$ , but they are deleted after merging.

Example 1 illustrates that the merging methods based on t-conorms are too cautious when most of the formulas are not involved in conflict while the renormalization based merging method may delete too much original information from the resulting knowledge base. According to [7], when knowledge bases are consistent with each other, it is preferable to use a t-norm based merging method. The maximum based merging method is preferable to the minimum based merging method (or any other t-norm based merging method) only if the inconsistency degree of  $B_1 \cup \dots \cup B_n$  is 1; that is, if there is a strong conflict among the sources of information.

### 3.1 Adaptive Combination Rules

Because of the pros and cons of the t-norm and t-conorm operators, it is not advisable to use only one of them when information sources partially agree with each other and only some of them are reliable. Several adaptive merging rules have been proposed to integrate both the t-norm and t-conorm operators.

**Conflict-respectful combination rule:** the normalization rules resolve the conflict between sources. However, as pointed out in [8], they may be very sensitive to rather small variations of possibility degrees around 0. In other word, the rules are not continuous in the vicinity of the total conflict expressed by  $h(\pi_{\oplus}) = 0$ . An adaptive combination rule was proposed in [10] to discount the result given by normalization rule by the inconsistency degree of the conjunctively merged knowledge bases, i.e.  $1 - h(\pi_{B_{\oplus}})$ . That is, we have the following modified conjunctive combination rule:  $\forall \omega$ ,

$$(\text{CoR-N}) \pi_{\oplus, \text{CoR-N}}(\omega) = \max(\pi_{N, \oplus}(\omega), 1 - h(\pi_{\oplus})).$$

**An adaptive rule in [8]** was proposed which considered  $j$  sources out of all the sources, where it was assumed that these  $j$  sources are reliable. Since it was not known which  $j$  sources were reliable, all the subsets with cardinality  $j$  were considered. The intermediary conjunctively merged results are then merged disjunctively. Given a possibilistic profile  $\mathcal{E} = \{B_1, \dots, B_N\}$  with  $\pi_i$  being the possibility distribution of  $B_i$ , the adaptive rule is defined as

$$\pi_{(j)}(\omega) = \max_{J \subseteq \mathbf{N}, |J|=j} \{\min_{i \in J} \{\pi_i(\omega) | \omega \in \Omega\}\}, \quad (7)$$

where  $\mathbf{N} = \{1, \dots, N\}$ .

A method to decide the value of  $j$  was given in [10]: let

$$m = \max\{|T| : h(T) = 1\}, \quad (8)$$

$$n = \max\{|T| : h(T) > 0\}, \quad (9)$$

where  $T \subseteq \mathcal{E}$  and  $h(T) = \max_{\omega} \min_{B_i \in T} \pi_i(\omega)$ , then,  $j$  is defined as  $m$  and  $N$  is defined as  $n$ , where  $n$  indicates that these  $n$  sources at least partially consistent and among them  $j$  sources are completely consistent.

**Another adaptive rule in [12]:** the rule proposed in [12] utilizes the maximum and the minimum operators. This operator is extended to more than two sources in [10] based on the adaptive rule in Equation 7. It is defined as follows:

$$\pi_{AD}(\omega) = \max\left(\frac{\pi_{(n)}(\omega)}{h(n)}, \min(\pi_{(m)}(\omega), 1 - h(n))\right),$$

where  $h(n) = \max\{h(T) | |T| = n\}$  as defined previously. This operator utilizes both  $\pi_{(n)}$  and  $\pi_{(m)}$ . It first *renormalizes*  $\pi_{(n)}$  which is the result of conjunctive merge of  $n$  sources and then applies the maximum operator to the conjunctively merged results of all subgroups with cardinality  $m$ , before combines the two parts using the maximum operator. So it is more adaptive and context dependent than adaptive rule in Equation 7.

**MCS-based adaptive merging in [11]:** an adaptive operator based on maximal consistent subsets (MCS) of  $\mathcal{E}$  was proposed in [11]. Suppose  $\mathcal{E}_1, \dots, \mathcal{E}_k$  are

all the maximal consistent subsets of  $\mathcal{E}$ , then the MCS-based operator is defined as

$$\pi_{MCS}(\omega) = \max_{i=1,\dots,k} \min_{B_j \in \mathcal{E}_i} \pi_{B_j}(\omega)$$

**Split-Combination merging approach in [26]:** The general idea of the Split-Combination (S-C) approach can be described as follows. Given a set of PKBs  $B_i$ , where  $i = 1, \dots, n$ , in the first step, we split them into  $B_i = \langle C_i, D_i \rangle$  with regard to a splitting method. In the second step, we combine all  $C_i$  by a t-conorm operator (the result is a PKB  $C$ ) and combine all  $D_i$  by a t-norm operator (the result is a PKB  $D$ ). The final result of the S-C combination method, denoted by  $B_{S-C}$ , is  $C \cup D$ . Different S-C methods can be developed by incorporating different ways of splitting the knowledge bases, while retaining the general S-C approach. Two different splitting methods have been given. One is called the Incremental splitting (I-S) method and the other is called the *free-formula based* splitting (F-S) method. The I-S method first searching for the splitting value using an incremental algorithm and then splits each PKB  $B_i$  into two subbases using the splitting value. The F-S method splits a PKB  $B$  into two subbases such that one of them contains formulae which are not in conflict in  $B$  and the other contains formulae which are in conflict. We call the merging operator based on the *I-S* method as Incremental Split-Combination (*I-S-C*) merging operator and the merging operator based on the free-formula based method as free-formula based split-combination (*F-S-C*) merging operator. Note that the *F-S-C* merging operator does not have a semantic counterpart.

In next section, we propose three evaluation criteria and compare existing merging operators in possibilistic logic with respect to them. As for merging operators in classical logic, we use  $\Delta_X$  to denote the merging operator  $X$ . For example,  $\Delta_{ct}$  is the t-conorm based merging operator.

## 4 Evaluation Criteria

### 4.1 Discriminating Power

Information is generally hard to obtain. Therefore, a good merging operator should preserve as much original information as possible. Inferential power is an important factor for evaluating a merging operator. Given two merging operators, it is natural to prefer the one leading to a merged base which can non-trivially infer more information. However, most merging operators can not be compared with regard to it. Therefore, we propose another evaluation method which is based on the *measure of non-specificity*.

In [17], a measure of possibilistic uncertainty, called *nonspecificity*, was proposed to generalize the Hartley measure of information [16]. Given a possibility distribution  $\pi$  on  $\Omega = \{\omega_1, \dots, \omega_n\}$ ,  $\pi(\omega_i)$  ( $i = 1, \dots, n$ ) are reordered as  $\pi_1 = l \geq \pi_2 \geq \dots \geq \pi_n$ , where  $l$  may be less than 1 (in that case,  $\pi$  is not normal). Let  $\pi_{i+1} = 0$ . The measure of nonspecificity of  $\pi$  is

$$H(\pi) = \frac{1}{l} \sum_{j=1}^n (\pi_j - \pi_{j+1}) \log_2 j \quad (10)$$

Given two PKBs  $B_1$  and  $B_2$ , we say the quality of  $B_1$  is better than that of  $B_2$  if  $H(\pi_{B_1}) < H(\pi_{B_2})$ , where  $\pi_{B_i}$  ( $i = 1, 2$ ) are possibility distributions of  $B_i$ .

Nonspecificity measures information carried by a possibility distribution. Now an interesting question is whether a knowledge base with stronger inferential power also has better quality. The answer is positive according to the following proposition.

**Proposition 2.**<sup>1</sup> *Given two consistent PKBs  $B_1$  and  $B_2$ ,  $\pi_{B_1}$  and  $\pi_{B_2}$  are possibility distributions associated with them. If  $B_1 \vdash_\pi (\phi, a)$  for any  $(\phi, a) \in B_2$ , then  $H(\pi_{B_1}) \leq H(\pi_{B_2})$ .*

Proposition 2 agrees with intuition that stronger inferential power means more information. Therefore, nonspecificity measures the discriminating power of a merging operator.

In the following, we define an ordering to compare two possibilistic merging operators with respect to their discriminating power.

**Definition 2.** *Let  $\Delta_1$  and  $\Delta_2$  be two merging operators in possibilistic logic. An ordering between  $\Delta_1$  and  $\Delta_2$ , denoted  $\preceq_{DP}$ , is defined as:*

$$\Delta_1 \preceq_{DP} \Delta_2 \quad \text{iff} \quad \forall \mathcal{E}, H(\pi_{\Delta_1(\mathcal{E})}) \leq H(\pi_{\Delta_2(\mathcal{E})}),$$

where  $\pi_{\Delta_i(\mathcal{E})}$  ( $i = 1, 2$ ) are the possibility distributions associated with  $\Delta_i(\mathcal{E})$ . It is clear that  $\preceq_{DP}$  is a partial preorder, i.e. it is reflective and transitive. As usual, we define the strict order  $\Delta_1 \prec_{DP} \Delta_2$  iff  $\Delta_1 \preceq_{DP} \Delta_2$  and  $\Delta_2 \not\preceq_{DP} \Delta_1$ . If  $\Delta_1 \prec_{DP} \Delta_2$ , then we say that  $\Delta_1$  is stronger than  $\Delta_2$  w.r.t their discriminating power.

We compare existing merging operators in possibilistic logic w.r.t their discriminating power.

**Proposition 3.** *When  $\mathcal{E}$  is consistent, then  $\Delta_{N,tn} \prec_{DP} \Delta_{ct}$ . However, this relationship does not hold when  $\mathcal{E}$  is inconsistent.*

**Proposition 4.** *The normalized conjunctive merging operator has stronger discriminating power than the conflict-respectful merging operator, i.e.,  $\Delta_{N,tn} \prec_{DP} \Delta_{C \circ R - N}$ .*

**Proposition 5.** *Both the  $j$ -source and MCS based adaptive merging operators have stronger discriminating power than the  $t$ -conorm based one. That is,  $\Delta_{\{j\}} \prec_{DP} \Delta_{ct}$  and  $\Delta_{MCS} \prec_{DP} \Delta_{ct}$ .*

**Proposition 6.** [26] *Let  $\mathcal{E} = \{B_1, \dots, B_n\}$  be a set of  $n$  PKBs. Let  $\Delta_{I-S-C}$ ,  $\Delta_{F-S-C}$  and  $\Delta_{ct}$  be the  $I$ - $S$ - $C$  operator,  $F$ - $S$ - $C$  operator and  $t$ -conorm based merging operator respectively. We have  $\Delta_{I-S-C} \prec_{DP} \Delta_{ct}$  and  $\Delta_{F-S-C} \prec_{DP} \Delta_{ct}$ .*

---

<sup>1</sup> All proofs of propositions can be found in a PhD thesis which is available at <http://gqi.limewebs.com/thesis.pdf>.



It has been shown in [26] that *I-S-C* merging operator and normalized conjunctive merging operator are not comparable *w.r.t* the discriminating power. However, we have the following proposition.

**Proposition 7.** [26] *Let  $B_1$  and  $B_2$  be two PKBs. Suppose the t-norm  $tn$  and t-conorm  $ct$  which are used to define the I-S-C operator is the minimum and an arbitrary t-conorm respectively. Let  $\gamma$  be the splitting point for the I-S-C operator which is obtained by the splitting algorithm. Suppose  $\gamma = Inc(B_1 \cup B_2)$ , then  $B_{min, N2} \subseteq B_{I-S-C}$ , but not vice versa.*

## 4.2 Computational Complexity

Computational complexity has been adopted as an important criterion to evaluate a solution in many AI problems, such as belief revision [15]. From an application point of view, computational efficiency is an important requirement when one selects a merging operator. It has been shown in [22] that in the propositional setting, merging is generally a hard task in the worst case. It is clear that computationally more efficient operators are preferable to more complex ones.

We assume that the reader is familiar with computational complexity (more details can be found in [18]) and we consider the following classes located at the first or the second level of polynomial hierarchy.

- $\Delta_2^p = P^{NP}$  is the class of decision problems solvable in polynomial time by a deterministic Turing machine equipped with an NP oracle.
- $\Sigma_2^p$  is the class of decision problems solvable in polynomial time by a non-deterministic Turing machine equipped with an NP oracle.
- $\Pi_2^p = co-\Sigma_2^p$ , where  $co-\Sigma_2^p$  is the class of problems whose answer is always the complement of those in  $\Sigma_2^p$ .

Note that the following relations have been conjectured in complexity theory:

$$P \subset NP, P \subset co-NP, NP \neq co-NP, NP \subset \Delta_2^p, co-NP \subset \Delta_2^p, co-NP \subset \Pi_2^p, \Delta_2^p = P^{NP} \subset \Sigma_2^p$$

In this paper, we consider *function problems*, problems that require an answer more elaborate than “yes” or “no”. Suppose  $X$  is a class of decision problems, the function problem associated with  $X$  is denoted by  $FX$ . For example,  $F\Delta_2^p$  is the set of problems solvable in polynomial time by a machine with access to an oracle for an NP problem.

In the following, we analyze the computational complexity of existing merging operators.

**Proposition 8.** *Generating a consistent PKB by a t-conorm based merging operator is in FP and generating a consistent PKB by a normalized conjunctive merging operator is in  $F\Delta_2^p$ .*

By Proposition 8, the t-conorm based merging operator is tractable. However, since it is a cautious operator that may drop too much information, we do not advocate using it unless the sources are strongly in conflict. Therefore, before

applying the t-conorm based merging operator, we should at least check the consistency of the union of original sources, which is a NP-complete task.

The computational complexity of the conflict-respectful operator is not harder than that of the normalized conjunctive merging operator.

**Proposition 9.** *Generating a consistent PKB by a conflict-respectful merging operator is in  $F\Delta_2^P$ .*

**Proposition 10.** *Suppose  $j = m$ , where  $m$  is defined by Equation (8), then generating a consistent PKB by the  $j$ -source based merging operator is in  $F\Delta_2^P$ . Generating a consistent PKB by the adaptive merging operator  $\Delta_{AD}$  is in  $F\Delta_2^P$ .*

The MCS-based merging operator is computationally harder than other operators because it needs to compute all the maximal consistent subbases.

**Proposition 11.** *Generating a consistent PKB by the MCS-based adaptive merging operator is  $F\Pi_2^P$ -hard.*

We now consider the computational complexity results given in [26].

**Proposition 12.** *Generating a consistent PKB using the I-S-C operator and F-S-C operator are in  $F\Delta_2^P$  ( $\mathcal{O}(n)$ ) and  $F\Pi_2^P$  respectively.*

According to above propositions, the t-conorm based merging operator is computationally easier than other merging operators. The computational complexities of most of merging operators in possibilistic logic remain at the first level of polynomial hierarchy, a low level of polynomial hierarchy. Finally, the MCS-based merging operator and F-S-C merging operator are computationally harder than other operators, i.e. their computational complexities are located at least at the second level of polynomial hierarchy.

### 4.3 Compatibility with Propositional Merging Operators

It has been pointed out in [14] that when the necessity degrees of all the possibilistic formulas are taken as 1, possibilistic logic is reduced to classical logic. So classical logic is a special case of possibilistic logic in which all the formulas have the same level of priority. Therefore, merging methods in possibilistic logic can be directly applied to merge classical knowledge bases. When comparing two different operators which are not comparable with regard to other criteria, we prefer possibilistic merging operators which are well-behaved in classical logic to those which are not. In classical logic, the main criterion for comparing merging operator is the rationality properties. In this section, we analyze the logical properties of different possibilistic merging operators in the flat case.

We first introduce the postulates for characterizing a propositional merging operator proposed in [19].

**Definition 3.** *Let  $\Delta$  be a propositional operator which assigns to a set of knowledge bases  $E$  a knowledge base  $\Delta(E)$ . Let  $E_1$  and  $E_2$  be two sets of knowledge bases,  $K$  and  $K'$  be two knowledge bases.  $\Delta$  is a propositional merging operator iff it satisfies the following postulates:*

(A1)  $\Delta(E)$  is consistent.

(A2) If  $E$  is consistent, then  $\Delta(E) \equiv \bigwedge E$ , where  $\bigwedge E = \bigwedge_{K_i \in E} K_i$ .

(A3) If  $E_1 \equiv E_2$ , then  $\Delta(E_1) \equiv \Delta(E_2)$ .

(A4) If  $K \wedge K'$  is not consistent, then  $\Delta(\{K\} \sqcup \{K'\}) \not\equiv K$ .

(A5)  $\Delta(E_1) \wedge \Delta(E_2) \models \Delta(E_1 \sqcup E_2)$ .

(A6) If  $\Delta(E_1) \wedge \Delta(E_2)$  is consistent, then  $\Delta(E_1 \sqcup E_2) \models \Delta(E_1) \wedge \Delta(E_2)$ .

**Proposition 13.** Let  $E = \{K_1, \dots, K_n\}$  be a set of knowledge bases. Let  $\Delta_{ct}(E)$  be the resulting knowledge base of a  $t$ -conorm based operator. Then  $\Delta_{ct}(E) \equiv \{\bigvee_{i=1}^n \phi_i : \phi_i \in K_i, i = 1, \dots, n\}$ . Let  $\Delta_{N,tn}(E)$  be a normalized  $t$ -norm operator. Then  $\Delta_{N,tn}(E) \equiv \top$ .

According to Proposition 13, the normalized conjunctive operators cannot be directly applied to merging classical knowledge bases because all the information is deleted after merging, whilst the  $t$ -conorm based operators take the disjunction of all the knowledge bases as the result of merging.

The logical properties of the operator  $\Delta_{ct}$  in the flat case are analyzed as follows.

**Proposition 14.** The  $t$ -conorm based operator  $\Delta_{ct}$  satisfies (A1), (A3), (A4) and (A5). It does not satisfy other postulates in general.

Next, we consider the adaptive merging operators.

First, the conflict-respectful merging operators are based on the normalized conjunctive operators. So, they cannot be directly applied to merge classical knowledge base.

The  $j$ -source based operator takes the disjunction of all the knowledge bases which are conjunction of  $j$  original knowledge bases as the result of merging, that is, we have  $\Delta_{\{j\}}(E) = \bigvee_{J \subseteq \mathbf{N}, |J|=j} \{\bigwedge_{i \in J} K_i\}$ .

**Proposition 15.** When  $j = m$ , where  $m$  is defined by Equation (8), then the  $j$ -source based operator  $\Delta_{\{j\}}$  satisfies (A1), (A2), (A3), (A4). It does not satisfies (A5) and (A6) in general.

According to Proposition 15, when  $j = m$ , the  $j$ -source based operator has good logical properties in the flat case.

The adaptive operator  $\Delta_{AD}$  is based on the  $j$ -source based operator. In the flat case, it is equivalent to the  $m$ -source based operator according to the following proposition:

**Proposition 16.** The merged result of the adaptive operator  $\Delta_{AD}$  is equivalent to  $\Delta_{\{m\}}(E)$ , i.e.,  $\Delta_{AD}(E) \equiv \Delta_{\{m\}}(E)$ .

The merged result of the MCS-based adaptive operator is the disjunctive of those knowledge bases which are conjunctions of maximal consistent subsets of  $E$ . That is, let  $\text{MAXCONS}(E)$  be the set of maximal consistent subsets of  $E$ , we have

$$\Delta_{MCS}(E) = \bigvee_{E_i \in \text{MAXCONS}(E)} \bigwedge_{K_{ij} \in E_i} K_{ij}.$$

The operator  $\Delta_{MCS}(E)$  is a commonly used syntax-based merging operator in the propositional setting [20].

**Proposition 17.** *The MCS-based adaptive operator  $\Delta_{MCS}$  satisfies (A1), (A2), (A3), (A4), (A5). It does not satisfy (A6) in general.*

In [26], it has been shown that, in the flat case, the *I-S-C* operator is reduced to the t-conorm based operator. We also have the following results for *F-S-C* operator [26].

**Proposition 18.** *The F-S-C merging operator  $\Delta_{F-S-C}$  satisfies (A1), (A2), (A4) and (A5). However, it does not satisfy (A3) and (A6) in general.*

According to the above propositions, the operators  $\Delta_{N,tn}$  and  $\Delta_{CoR-N}$  cannot be applied to merging classical knowledge bases. The operator  $\Delta_{AD}$  and the operator  $\Delta_{\{m\}}$  are equivalent in the flat case. The MCS-based adaptive operator satisfies more postulates than all the other operators in possibilistic logic in the flat case.

## 5 Conclusions

In this paper, we first gave a belief survey of two important classes of merging operators in possibilistic logic: semantic merging operators and adaptive merging operators. We then proposed several evaluation criteria to compare these merging operators: discriminating power, computational complexity and compatibility with propositional merging. The comparison results will be useful for users to select appropriate merging operator(s) for specific applications.

## References

1. Amgoud, L., Kaci, S.: An argumentation framework for merging conflicting knowledge bases. *IEEE Transactions on Knowledge and Data Engineering* 3(2), 208–220 (2007)
2. Benferhat, S., Cayrol, C., Dubois, D., Lang, L., Prade, H.: Inconsistency management and prioritized syntax-based entailment. In: *Proc. of IJCAI 1993*, pp. 640–645 (1993)
3. Benferhat, S., Dubois, D., Prade, H.: From semantic to syntactic approaches to information combination in possibilistic logic. In: *Aggregation and Fusion of Imperfect Information*, pp. 141–151. Physica Verlag, Heidelberg (1997)
4. Benferhat, S., Dubois, D., Kaci, S., Prade, H.: Possibilistic merging and distance-based fusion of propositional information. *Annals of Mathematics and Artificial Intelligence* 34, 217–252 (2002)
5. Benferhat, S., Kaci, S.: Fusion of possibilistic knowledge bases from a postulate point of view. *Int. J. Approx. Reasoning* 33(3), 255–285 (2003)
6. Benferhat, S., Dubois, D., Prade, H., Williams, M.-A.: A practical approach to fusing prioritized knowledge bases. In: Barahona, P., Alferes, J.J. (eds.) *EPIA 1999. LNCS (LNAI)*, vol. 1695, pp. 223–236. Springer, Heidelberg (1999)

7. Benferhat, S., Sossai, C.: Reasoning with multiple-source information in a possibilistic logic framework. *Information Fusion* 7(1), 80–96 (2006)
8. Dubois, D., Prade, H.: Representation and combination of uncertainty with belief functions and possibility measures. *Computational Intelligence* 4, 244–264 (1988)
9. Dubois, D., Prade, H., Testemale, C.: Weighted fuzzy pattern matching. *Fuzzy Sets and Systems* 28, 313–331 (1988)
10. Dubois, D., Prade, H.: Possibility theory and data fusion in poorly informed environments. *Control Engineering Practice* 2(5), 811–823 (1994)
11. Dubois, D., Fargier, H., Prade, H.: Multiple source information fusion: a practical inconsistency tolerant approach. In: *Proc. of IPMU 2000*, pp. 1047–1054 (2000)
12. Dubois, D., Prade, H.: Combination of fuzzy information in the framework of possibility theory. In: Abidi, M.A., Gonzalez, R.C. (eds.) *Data Fusion in Robotics and Machine Intelligence*, pp. 481–505 (1992)
13. Dubois, D., Prade, H.: Possibility theory and data fusion in poorly informed environments. *Control Engineering Practice* 2(5), 811–823 (1994)
14. Dubois, D., Lang, J., Prade, H.: Possibilistic logic. In: *Handbook of Logic in Artificial Intelligence and Logic Programming*, vol. 3, pp. 439–513. Oxford University Press, Oxford (1994)
15. Eiter, T., Gottlob, G.: On the complexity of propositional knowledge base revision, updates, and counterfactuals. *Artificial Intelligence* 57, 227–270 (1992)
16. Hartley, R.: Transmission of information. *Bell System Technical Journal* 7, 535–563 (1928)
17. Higashi, M., Klir, G.: Measures of uncertainty and information based on possibility distributions. *International Journal of General Systems* 9(1), 43–58 (1983)
18. Johnson, D.S.: A catalog of complexity classes. In: van Leeuwen, J. (ed.) *Handbook of Theoretical Computer Science*, pp. 67–161 (1990)
19. Konieczny, S., Pino Pérez, R.: On the logic of merging. In: *Proc. of KR 1998*, pp. 488–498 (1998)
20. Konieczny, S.: On the difference between merging knowledge bases and combining them. In: *Proc. of KR 2000*, pp. 135–144 (2000)
21. Konieczny, S., Pino Pérez, R.: Merging information under constraints: A logical framework. *Journal of Logic and Computation* 12(5), 773–808 (2002)
22. Konieczny, S., Lang, J., Marquis, P.:  $DA^2$  merging operators. *Artificial Intelligence* 157(1-2), 49–79 (2004)
23. Liberatore, P., Schaerf, M.: Arbitration (or How to Merge Knowledge Bases). *IEEE Transactions on Knowledge and Data Engineering* 10(1), 76–90 (1998)
24. Qi, G., Liu, W., Glass, D.H.: Combining individually inconsistent prioritized knowledge bases. In: *Proc. of NMR 2004*, pp. 342–349 (2004)
25. Qi, G., Liu, W., Bell, D.A.: Combining multiple knowledge bases by negotiation: A possibilistic approach. In: Godo, L. (ed.) *ECSQARU 2005. LNCS (LNAI)*, vol. 3571, pp. 501–513. Springer, Heidelberg (2005)
26. Qi, G., Liu, W., Glass, D.H., Bell, D.A.: A split-combination approach for merging possibilistic knowledge bases. *Annals of Mathematics and Artificial Intelligence* 48(1-2), 45–84 (2006)

# Modelling and Reasoning in Metamodelling Enabled Ontologies<sup>\*</sup>

Nophadol Jekjantuk<sup>1</sup>, Gerd Gröner<sup>2</sup>, and Jeff. Z. Pan<sup>1</sup>

<sup>1</sup> University of Aberdeen, United Kingdom

<sup>2</sup> University of Koblenz-Landau, Germany

**Abstract.** Ontologies are expected to play an important role in many application domains, as well as in software engineering in general. One problem with using ontologies within software engineering is that while UML, a widely used standard for specifying and constructing the models for a software-intensive system, has a four-layer metamodelling architecture, the standard Web Ontology Language (OWL) does not support reasoning over layered metamodels. OWL 2 provides simple metamodelling by using a punning approach, however, the interpretation function is different based on the context, which leads to non-intuitive results. The OWL FA Language has a well defined metamodelling architecture. However, there is no study and tool for support reasoning over OWL FA. In this paper, we discuss some reasoning tasks in OWL FA. We also introduce the OWL FA Tool kit, a simple tool kit for manipulating and reasoning with OWL FA.

## 1 Introduction

Metamodelling appeals in many applications areas (such as UML [11], Model Driven Architecture [2], XML [13] and E-Commerce). It is not only the underpinning of modelling languages such as UML, but also central to OMG's MDA-based computing.

The W3C Web Ontology Language (OWL) [10] in combination with reasoning is already used in various other research areas like in model-driven software engineering in order to exploit the expressiveness of OWL and the usage of inference. However, the lack of a formal OWL language or OWL extension which supports metamodelling is an obstacle for the usage of OWL in other complex application areas.

The Resource Description Framework (RDF) and OWL Full support metamodelling by allowing users to use the built-in vocabulary without restrictions, which introduces an undecidability problem. OWL [10] provides formal semantics focused on conceptual modelling and adaptability of inference using DL reasoners and reasoning algorithms, but OWL does not support layered reasoning. OWL 2 provides simple metamodelling with semantics which correspond to

---

<sup>\*</sup> This paper is an extended version of [5] which was presented at the International workshop on OWL: Experience and Directions (OWL-ED2009).

the contextual semantics defined in [6], however, it has been shown in [9] that these can lead to non-intuitive results.

For example, the following axioms state that *Eagle* is an *Endangered* species, and that *Harry* is an *Eagle*:

$$\text{ClassAssertion}(\text{Endangered Eagle}) \quad (1)$$

$$\text{ClassAssertion}(\text{Eagle Harry}) \quad (2)$$

The axioms 1, 2 could be interpreted by DL reasoner as follows:

$$\text{ClassAssertion}(\text{Cls} - \text{Endangered Ind} - \text{Eagle}) \quad (3)$$

$$\text{ClassAssertion}(\text{Cls} - \text{Eagle Ind} - \text{Harry}) \quad (4)$$

The names of concepts and individuals do not interact with each other even they are sharing the same name, e.g. *Eagle* is represented as individual by the name *Ind - Eagle* and as class by the name *Cls - Eagle*. This kind of metamodelling is often referred to as punning. Let us consider the following axioms:

$$\text{SameIndividuals}(\text{Aquila Eagle}) \quad (5)$$

$$\text{ClassAssertion}(\text{not}(\text{Aquila}) \text{ Harry}) \quad (6)$$

The axioms 5, 6 could be safely added to the ontology in contextual semantics, but under layered semantics this ontology is inconsistent because 5 indicates the meta-individual equality since the axiom  $\text{Eagle} \approx \text{Aquila}$  indicates the equivalence of the two concepts *Eagle* and *Aquila*. However, axiom 6 describes that *Harry* is not in *Aquila* which leads to the contradiction in combination with axiom 4.

In this paper, we present modelling and reasoning algorithms for OWL FA knowledge bases. For the reasoning service, an OWL FA ontology is transformed to a set of OWL DL ontologies, then existing DL reasoners are applied to the transformed knowledge base. The syntax and semantics of OWL FA is described in Section 2. Modelling in OWL FA is demonstrated in Section 3. In Section 4 reasoning in OWL FA is described. This contains a reduction to OWL DL knowledge bases and reasoning algorithms in order to propagate conditions between different modelling layers. Features of OWL FA Tool Kit are detailed in Section 5. The early evaluation are presented in Section 6. Then, related work and direction of OWL FA are discussed in Section 7 and Section 8 respectively.

## 2 OWL FA Syntax and Semantics

OWL FA [9] enables metamodelling. It is an extension of OWL DL, which refers to the description logic  $\mathcal{SHOIN}(\mathcal{D})$ . Ontologies in OWL FA are represented in a layered architecture. This architecture is mainly based on the architecture of RDFS(FA) [8].

OWL FA specifies a layer number in class constructors and axioms to indicate the layer they belong to.

Let  $CN \in \mathbf{V}_{C_i}$  be an atomic class name in layer  $i$  ( $i \geq 0$ ),  $R$  an OWL FA-property in layer  $i$ ,  $o \in \mathbf{I}$  an individual,  $T \in \mathbf{V}_{DP}$  a datatype property name, and  $C, D$  OWL FA-classes in layer  $i$ . Valid OWL FA-classes are defined by the abstract syntax:

$$C ::= \top_i \mid \perp \mid CN \mid \neg_i C \mid C \sqcap_i D \mid C \sqcup_i D \mid \{o\} \mid \exists_i R.C \mid \\ \mid \forall_i R.C \mid \leq_i nR \mid \geq_i nR \mid \\ (\text{if } i = 1) \exists_1 T.d \mid \forall_1 T.d \mid \leq_1 nT \mid \geq_1 nT$$

The semantics of OWL FA is a model theoretic semantics, which is defined in terms of interpretations. In other words, the semantics of two layers which can be considered as TBox and ABox are same as in OWL DL. The idea of OWL FA is that the interpretation depends on the layer but is still an OWL DL interpretation. Given an OWL FA alphabet  $\mathbf{V}$ , a set of built-in datatype names  $\mathbf{B} \subseteq \mathbf{V}_D$  and an integer  $k \geq 1$ , an *OWL FA interpretation* is a tuple of the form  $\mathcal{J} = (\Delta^{\mathcal{J}}, \cdot^{\mathcal{J}})$ , where  $\Delta^{\mathcal{J}}$  is the domain (a non-empty set) and  $\cdot^{\mathcal{J}}$  is the interpretation. In the rest of the paper, we assume that  $i$  is an integer such that  $1 \leq i \leq k$ . The interpretation function can be extended to give semantics to OWL FA-properties and OWL FA-classes. Let  $RN \in \mathbf{V}_{AP_i}$  be an abstract property name in layer  $i$  and  $R$  be an abstract property in layer  $i$ . Valid OWL FA abstract properties are defined by the abstract syntax:  $R ::= RN \mid R^-$ , where for some  $x, y \in \Delta_{A_{i-1}}^{\mathcal{J}}$ ,  $\langle x, y \rangle \in R^{\mathcal{J}}$  iff  $\langle y, x \rangle \in R^{-\mathcal{J}}$ . Valid OWL FA datatype properties are datatype property names. The interpretation function is explained in detail in [9].

An interpretation  $\mathcal{J}$  satisfies an ontology  $\Sigma$  if it satisfies all the axioms in  $\Sigma$ .  $\Sigma$  is *satisfiable* (*unsatisfiable*) iff there exists (does not exist) such an interpretation  $\mathcal{J}$  that satisfies  $\Sigma$ . Let  $C, D$  be OWL FA-classes in layer  $i$ ,  $C$  is *satisfiable* w.r.t.  $\Sigma$  iff there exists an interpretation  $\mathcal{J}$  of  $\Sigma$  s.t.  $C^{\mathcal{J}} \neq \emptyset_i$ ;  $C$  subsumes  $D$  w.r.t.  $\Sigma$  iff for every interpretation  $\mathcal{J}$  of  $\Sigma$  we have  $C^{\mathcal{J}} \subseteq D^{\mathcal{J}}$ .

### 3 Modelling of Metamodelling Enabled Ontologies

In this section, we present the way to express metamodelling enabled ontologies in OWL 2. The layer information is encapsulated in custom annotation property called "Layer". This is different from [5] because we realise that creating a new syntax for OWL FA is unnecessary since we could store layer information as annotation properties. Moreover, this ontology conforms still to the OWL 2 syntax like the punning style which is another way to capture a simple modelling in OWL 2. Although, the layer numbers can/should be encapsulated by tools, there are two rules of thumb to help users to get the number right. Firstly, the subscript numbers are only used to indicate a sub-ontology (e.g.  $\mathcal{O}_2$ ), a constructor (e.g.  $\exists_2$ ), or axiom symbols (e.g.  $\sqsubseteq_2, :_2$ ) in a sub-ontology. Secondly, subscript numbers of the constructors and axiom symbols indicate the sub-ontology that the class descriptions constructed by these constructors and axioms belong to.

The following example shows how to model an Endangered Species ontology with the DL syntax and then convert it into OWL 2 functional syntax. The main



reason for using functional syntax is that it is obvious to see which layer they belong to.

*Example 1.* Endangered Species ontology expressed in DL syntax as follow:

$$\text{Eagle} :_2 \text{ Endangered} \quad (7)$$

$$\text{Aquila} :_2 \text{ Endangered} \quad (8)$$

$$\text{Aquila} \approx_2 \text{ Eagle} \quad (9)$$

$$\text{Eagle} \sqsubseteq_1 \text{ Bird} \quad (10)$$

$$\text{Aquila} \sqsubseteq_1 \text{ Bird} \quad (11)$$

$$\text{Harry} :_1 \text{ Eagle} \quad (12)$$

*Example 2.* Endangered Species ontology expressed in OWL 2 syntax as follow:

$$\text{ClassAssertion(Annotation(Layer"2")Endangered Eagle)} \quad (13)$$

$$\text{ClassAssertion(Annotation(Layer"2")Endangered Aquila)} \quad (14)$$

$$\text{SameIndividuals(Annotation(Layer"2")Eagle Aquila)} \quad (15)$$

$$\text{SubClassOf(Annotation(Layer"1")Eagle Bird)} \quad (16)$$

$$\text{SubClassOf(Annotation(Layer"1")Aquila Bird)} \quad (17)$$

$$\text{ClassAssertion(Annotation(Layer"1")Eagle Harry)} \quad (18)$$

## 4 Reasoning in OWL FA

Now we briefly discuss some reasoning tasks in OWL FA. According to the layered architecture, the knowledge base  $\Sigma$  in OWL FA is divided into a sequence of knowledge bases  $\Sigma = \Sigma_1, \dots, \Sigma_k$ , whereas  $k$  is the number of layers. Since individuals in layer  $i + 1$  can be classes and properties in layer  $i$ , this also affects the axioms of the layer below. Hence, individual axioms in the knowledge base  $\Sigma_{i+1}$  can be considered as class axioms in the knowledge base  $\Sigma_i$ .

In an OWL FA knowledge base  $\Sigma$ ,  $\Sigma_2, \dots, \Sigma_k$  are *SHIQ* knowledge bases, i.e. nominals are not allowed. A nominal in a higher layer can lead to unsatisfiability of the knowledge bases. An interesting feature of  $\Sigma$  is that there could be interactions between  $\Sigma_i$  and  $\Sigma_{i+1}$ .

### 4.1 Preprocessing

In this section, we discuss how to reduce the reasoning problem in OWL FA into a reasoning problem in OWL DL.

**Definition 1.** Let  $\Sigma = \langle \Sigma_1, \dots, \Sigma_k \rangle$  be an OWL FA knowledge base, where each of  $\Sigma_1, \dots, \Sigma_k$  is consistent.  $\Sigma^* = \langle \Sigma_1^*, \dots, \Sigma_k^* \rangle$ , called the explicit knowledge base, is constructed by making all the implicit atomic class axioms, atomic property axioms, individual equality axioms explicit.  $\diamond$

As we have a finite set of vocabulary, we have the following Lemma.

**Lemma 1.** *Given an OWL FA knowledge base  $\Sigma = \langle \Sigma_1, \dots, \Sigma_k \rangle$ . The explicit knowledge base  $\Sigma^*$  (OWL DL knowledge base) can be calculated from  $\Sigma$  in finite steps.*

*Proof.* When  $k = 1$ , we can calculate the explicit knowledge base  $\Sigma_1^*$  in finite steps because the sets of names of classes (in layer 1), roles (in layer 1) and individuals are finite. When  $k > 1$ , let us assume that we can calculate the explicit knowledge bases  $\Sigma'_1, \dots, \Sigma'_i$  (where  $1 \leq i < k$ ) from  $\Sigma_1, \dots, \Sigma_i$  in finite steps. We add all the class and property equality axioms in  $\Sigma'_i$  to  $\Sigma_{i+1}$ . If the updated  $\Sigma_{i+1}$  is consistent. Then, we can make the implicit individual equality axioms (if any) explicit and add new class and property equality axioms into  $\Sigma'_i$ . Thus, we can calculate  $\Sigma''_1, \dots, \Sigma''_i$  in finite steps. As the individual names in  $\Sigma_{i+1}$  are finite, we can calculate the explicit knowledge bases  $\Sigma_1^*, \dots, \Sigma_{i+1}^*$  in finite steps.

Note that if a class description is not defined in  $\Sigma_i$  (i.e., if it is not equivalent to any atomic class), it is not represented by any meta-individual in  $\Sigma_{i+1}$ . This suggests the connections between  $\Sigma_i$  and  $\Sigma_{i+1}$  are atomic classes and properties in  $\Sigma_i$ , which are meta-individuals in  $\Sigma_{i+1}$ .

We now present the algorithm **Reduce**, that will reduce an OWL FA knowledge base  $\Sigma$  into a set of OWL DL knowledge bases  $\langle \Sigma_1^*, \dots, \Sigma_k^* \rangle$ . This algorithm is based on Definition 1 and Lemma 1. The algorithm takes an OWL FA KB  $\Sigma$  as input and returns a set of OWL DL KB  $\langle \Sigma_1, \dots, \Sigma_k \rangle$ . The Algorithm **Reduce** is shown in Algorithm 1.

The following theorem shows the termination of the algorithm **Reduce**, applied to an OWL FA KB  $\Sigma$ .

**Theorem 1.** *Given an OWL FA knowledge base  $\Sigma = \langle \Sigma_1, \dots, \Sigma_k \rangle$ , then **Reduce** ( $\Sigma$ ) terminates.*

*Proof.* Termination of algorithm **Reduce** is straightforward from Lemma 1, which we can construct  $\langle \Sigma_1^*, \dots, \Sigma_k^* \rangle$  from  $\Sigma$  in finite step and a sets of class, property and individual equality axioms are finite. Thus, algorithm **Reduce** always terminates.

Here is the result from applying the Algorithm **Reduce** to the OWL FA KB  $\Sigma$ :

$$\Sigma_2 = \{ \text{Endangered} : \text{Eagle}, \text{Endangered} : \text{Aquila}, \text{Eagle} = \text{Aquila} \}$$

$$\Sigma_1 = \{ \text{Eagle} : \text{Harry}, \text{Eagle} \sqsubseteq \text{Bird}, \text{Aquila} \sqsubseteq \text{Bird}, \text{Aquila} \equiv \text{Eagle} \}$$

## 4.2 Consistency Checking

In this section, we present the algorithm **Consistent**, that will check the consistency of an OWL FA knowledge base  $\mathcal{O}$ . We can reduce an OWL FA knowledge base to a collection of OWL DL knowledge bases, therefore existing DL reasoner capabilities can be used. Consistency checking for OWL FA is done in two steps:

**Algorithm 1.** Reduce**Input:** OWL FA KB  $\Sigma$ **Output:** satisfiable and a set of OWL DL KB  $\langle \Sigma_1, \dots, \Sigma_k \rangle$ .

---

```

1: boolean satisfiable = true;
2: Collect axioms from the layer number and store in  $L_0, \dots, L_n$ 
3: Create knowledge base  $\Sigma_i = (L_0, L_1), \dots, \Sigma_k = (L_{n-1}, L_n)$ 
4: repeat
5:    $ce_i = \emptyset, pe_i = \emptyset$  and  $oe_i = \emptyset$ 
6:   Check consistency of  $\Sigma_i$  with DL reasoner
7:   if  $\Sigma_i$  is consistent then
8:     for each  $\Sigma_i^*$  ( $1 \leq i \leq k$ ) do
9:       Identify the new concept equality in  $\Sigma_i^*$  and store it in  $ce_i$ 
10:      Identify the new property equality in  $\Sigma_i^*$  and store it in  $pe_i$ 
11:      Identify the new individual equality in  $\Sigma_i^*$  and store it in  $oe_i$ 
12:      for each  $ce_i, pe_i$  and  $oe_i$  do
13:        if  $i = 1$  then
14:          Add  $ce_i$  and  $pe_i$  as individual equality into  $\Sigma_{i+1}^*$ 
15:        else if  $i = k$  then
16:          Add  $oe_i$  as property or concept equality into  $\Sigma_{i-1}^*$ 
17:        else
18:          Add  $ce_i$  and  $pe_i$  as individual equality into  $\Sigma_{i+1}^*$ 
19:          Add  $oe_i$  as property or concept equality into  $\Sigma_{i-1}^*$ 
20:        end if
21:      end for
22:    end for
23:  else
24:    satisfiable = false;
25:  end if
26: until ( $ce_i = \emptyset$  &  $pe_i = \emptyset$  &  $oe_i = \emptyset$ ) || satisfiable = false;
27: return  $\langle \Sigma_1, \dots, \Sigma_k \rangle$ .

```

---

First, we check the syntax of OWL FA. For example,  $\Sigma = \{C \sqsubseteq_2 D, C \sqsubseteq_3 E\}$  is non-well formed because in OWL FA we do not allow OWL class construct between layer except an instance-of relationship. Secondly, we check the consistency of each OWL DL-knowledge base that is computed from the OWL FA knowledge base with an existing DL reasoner. The Algorithm **Consistent** is shown in Algorithm 2.

We invite the reader to note that *check-dl-consistent* is a function call to a DL Reasoner.

**Theorem 2.** *Given an OWL FA knowledge base  $\Sigma = \langle \Sigma_1, \dots, \Sigma_k \rangle$ .  $\Sigma$  is consistent iff each  $\Sigma_i^*$  ( $1 \leq i \leq k$ ) is consistent.  $\diamond$*

Theorem 2 shows we can reduce the OWL FA-knowledge base consistency problem to the OWL DL-knowledge base consistency problem.

*Proof.* The consistency check of an OWL FA KB with  $\text{Consistent}(\Sigma)$  is straightforward from Lemma 1. We can construct  $\langle \Sigma_1^*, \dots, \Sigma_k^* \rangle$  from  $\Sigma$  in finite steps

---

**Algorithm 2.** Consistent

---

**Input:** OWL FA Knowledge Base  $\Sigma = \langle \Sigma_1, \dots, \Sigma_k \rangle$ **Output:** *true* if  $\Sigma$  is consistent, *false* otherwise

```

1: Ont =  $\emptyset$ 
2: Check OWL FA syntax
3: Ont = Reduce( $\Sigma$ );
4: for each  $\Sigma_i$  in Ont do
5:   check-dl-consistent( $\Sigma_i$ )
6:   if  $\Sigma_i$  is not consistent then
7:     Return false
8:   end if
9: end for
10: return true.

```

---

then we check the consistency for each  $\Sigma_i^*$  with a DL reasoner. Therefore, the OWL FA knowledge base  $\Sigma$  is consistent if and only if each  $\Sigma_i^*$  ( $1 \leq i \leq k$ ) is consistent.

Let us consider the following axiom by inserting it into OWL FA KB  $\Sigma$  (cf. example 2):

$$\text{ClassAssertion(Annotation(Layer"1")Not(Aquila) Harry)} \quad (19)$$

It is obvious to see that this axiom will make  $\Sigma_1^*$  inconsistent, which leads to an inconsistent of OWL FA KB  $\Sigma$  because the meta-individual equality axiom  $\text{Eagle} \approx_2 \text{Aquila}$  indicates the equivalence of the two concepts Eagle and Aquila, and  $\text{Harry}^{\mathcal{J}}$  cannot be both in and not in  $\text{Eagle}^{\mathcal{J}}$ .

### 4.3 Instance Retrieval

Instance retrieval in OWL FA is trivial because after the reduction process, we get a set of OWL 2 DL ontology then we could perform instance retrieval against those ontologies. However, without specifying a target ontology, it is not efficient since, we have to go through all ontologies in a set. Therefore, we need a smart algorithm for instance retrieval for OWL FA, in order to select the right ontology that contains a target concept. Firstly, we need to search for a target concept in each ontology of the set. This step does not require any DL reasoner. Then, we could perform instance retrieval against a selected ontology with a DL reasoner. A formal definition of instance retrieval for OWL FA is given in Definition 2.

**Definition 2.** *Given an ABox  $\mathcal{A}_i$  and a query  $Q$ , i.e., a concept expression, find all individuals  $a$  such that  $a$  is an instance of  $Q$ , i.e.,  $\{a \mid \forall a \in \mathcal{A}_i, a : Q\}$ .  $\diamond$*

We present the instance retrieval for OWL FA in Algorithm 3. The algorithm *instanceOf* will take an OWL FA ontology  $\mathcal{O}$  and a concept  $C$  as input. The algorithm returns a set containing the instances of concept  $C$ .

Due to space limitations, we cannot describe all reasoning tasks for OWL FA in this paper, however, since we can reduce OWL FA into a set of OWL DL ontologies then all existing DL reasoner's capabilities can be used.

---

**Algorithm 3.** instanceOf

---

**Input:** OWL FA Knowledge Base  $\Sigma = \langle \Sigma_1, \dots, \Sigma_k \rangle$  and a concept  $C$ **Output:** A set contains instance of concept  $C$ 

```

1: ind =  $\emptyset$ 
2: Ont =  $\emptyset$ 
3: Ont = Reduce( $\Sigma$ );
4: for each  $\Sigma_i$  in Ont do
5:   if  $\Sigma_i$  contains  $C$  then
6:     ind = get-dl-instance-of( $\Sigma_i, C$ )
7:   end if
8: end for
9: return ind.

```

---

## 5 OWL FA Tool Kit

In this section, we reintroduce the OWL FA Tool Kit, a simple graphic user interface for modeller to create an OWL FA ontology and perform reasoning over it. The OWL FA Tool Kit contains features as follows:

- Editor - for checking the OWL FA ontology before performing the reasoning.
- Ontology Consistency Checker- for checking whether a given metamodelling enabled ontology is consistent.
- Concept Satisfiability Checker - for verifying whether a concept  $A$  is a non-empty set in a given OWL FA ontology  $\mathcal{O}$ .
- Query Answering - for accessing information form a given metamodelling enabled ontologies by using SPARQL queries.
- Export a collection of OWL DL to files - for separating the domain knowledge from its meta knowledge.

## 6 Evaluation

In this section, we compare the metamodelling in OWL FA with OWL 2 as OWL 2 is the only OWL language that can support metamodelling and it has tools support.

### 6.1 Use Case 1: Consistency Checking

OWL 2 provides simple metamodelling with semantics which correspond to the contextual semantics defined in [6], however, it has been shown in [9] that these can lead to non-intuitive results.

Let us consider an ontology from Example 2 in Section 3 and the axioms 19. This ontology is consistent when we perform consistency checking with any existing DL reasoner. The existing DL reasoner does not take layer information which are described as annotation property into account and it interpret this ontology with contextual semantics.

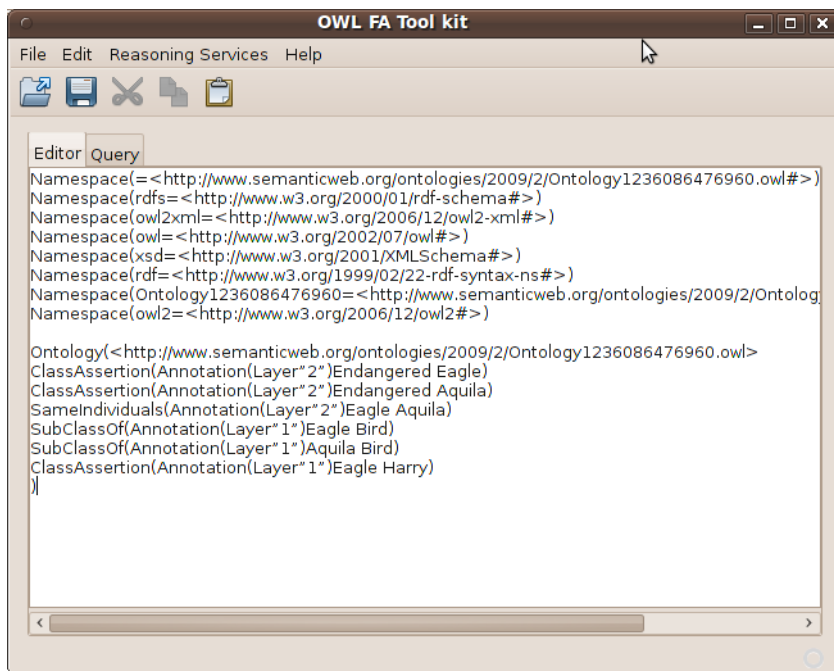


Fig. 1. OWL FA Tool Kit

This ontology is inconsistent based on layered architecture in OWL FA as we described in Section 4. OWL FA Tool Kit takes layer information into account and maintain a relationship between element that sharing the same URIs.

## 6.2 Use Case 2: Instance Retrieval

In OWL 2, if we do not provide the explicit instantiation between class and individual, it is difficult for any existing DL reasoner to discover those information because in contextual semantics, classes and individuals are interpreted independently. Let consider an ontology from Example 2 and remove axioms 14 and 15 from the ontology. Then, we would like to retrieve all objects that belong to *Endangered*. Without adding an axiom to indicate that *Aquila* is an *Endangered* then, *Aquila* is not included in the answer set. Although, concept *Aquila* is equivalent to concept *Eagle* but the interpretation of concept *Eagle* and object *Eagle* are independent from each other. Therefore, the existing DL reasoner could not found the relation between *Aquila* and *Endangered*.

For OWL FA and our tool kit, it returns more complete answer sets because in the reduction process, we propagate all class and property equalities to be object equalities in the higher layer and propagate all object equalities to be class or property equalities in the lower layer.

## 7 Related Work

OWL FA was introduced in [9] as a metamodelling extension of OWL. Motik [6] addressed metamodelling in OWL with two different semantics. The contextual semantics (or  $\pi$ -semantics) uses punning, i.e. names are replaced by distinct names for concepts, individuals and roles. This is like the different representation of an object in the OWL DL ontologies  $\Sigma_i$  in OWL FA. OWL2 [7] provides simple metamodelling features which is based on the contextual approach. The other semantics is the HiLog semantics (or  $\nu$ -semantics). The HiLog semantics is stronger than the  $\pi$ -semantics. The concepts and individual interpretations are not independent.

De Giacomina et al. [4] proposed the HiDL – Lite language, which adds one layer on top of the DL-Lite $\mathcal{R}$  language. This supports meta-classes and meta-properties and presents the query answering algorithm by reducing HiDL – Lite to DL-Lite $\mathcal{R}$  with the intention of using an existing DL-Lite reasoner. In OWL FA the semantics of meta-level are same as domain knowledge unlike HiDL-Lite that semantics of the meta-level need to re-define.

Description Logic reasoning is applied to UML models in [1,12,3]. The models are transformed into DL representations. Reasoning is used to check consistency of models and between models. However, metamodels are not considered.

## 8 Discussion

In this section, we discuss the future direction of OWL FA. Although OWL FA has a well defined metamodelling architecture, OWL does not support cross layer constraints. Let's take the well known Endangered species as an example. One would like to define a constraint on a meta-concept **Endangered** that all instances of these meta-concept have only 3000 individuals. Therefore, if a concept **Eagle** is an instance of the meta-concept **Endangered**, the constraint should be applied to the concept **Eagle** as well. We have an idea how to express this kind of constraint by using a `TopObjectProperty` in OWL 2. We can then express this **Endangered** requirement as  $\top \sqsubseteq \leq 3000 \sqcup .\text{Endangered}$  then propagate this constraint to all instances of **Endangered**. This is beyond OWL FA so we would like to investigate on enriching OWL FA toward OWL 2 FA. Another issue is that, in the cross layer constraints or restrictions for metamodelling in ontologies that we described in sections 2-5, we show that OWL FA is able to capture multiple layers better than OWL 2. However, the constraints or restrictions in OWL FA are relations between two layers. Let's consider an `ObjectProperty` assertion in layer  $M_2$ , **Endangered liveIn Continent**, which is expressed by the `ObjectProperty liveIn`. This constraint can only be used to validate the model between the layers  $M_2$  and  $M_1$ . We plan to investigate that it is possible to propagate the constraints across multiple layers. We are thinking about to use a meta prefix (`meta-`) like `meta_liveIn` that would still be an object property in the  $M_2$  layer and there the object property assertions in layer  $M_1$  remain unchanged like `liveIn(Eagle, Europe)`. However, we could also propagate this property assertion from model  $M_1$  semantically

to model  $M_0$ . For instance the property assertion `meta_leveln(Eagle, Europe)` in  $M_1$  becomes the subclass axiom `Eagle  $\sqsubseteq$   $\exists$  leveln.Europe` in the model  $M_0$ . This would be very interesting because we could specify all the constraints only in higher layers, then propagate them down to the lower layers automatically.

## 9 Conclusion

In this paper, we reintroduce OWL FA language and demonstrate how to model the metamodelling enabled ontology, followed by a description of reasoning in OWL FA for different reasoning tasks. And the reduction from an OWL FA knowledge base into OWL DL knowledge bases algorithms are describes. These algorithms use standard DL reasoning as a black-box service. Based on the given examples, a metamodelling enables ontology is described in OWL 2 DL.

We have shown that we can make use of the existing DL reasoners to reason over OWL FA knowledge base. As we discussed in section 4, we can calculate the explicit OWL DL knowledge base  $\Sigma_i^*$  from OWL FA knowledge base  $\Sigma^*$ .

We have implemented the OWL FA Tool Kit for modeller to manipulating and reasoning over OWL FA standard and plan to incorporate these into the TrOWL<sup>1</sup> reasoning infrastructure.

In the future, we would like to enrich OWL FA language toward the direction we described in discussion section in order to increase expressive power of the language such as propagate constraints between layers. Moreover, we would like to apply the fixed-layer architecture to OWL 2 DL which has more expressive power than OWL DL. And we plan to provide tools along with a reasoning mechanism for OWL 2 FA.

## Acknowledgements

This work has been partially supported by the European Project Marrying Ontologies and Software Technologies (MOST ICT 2008-216691).

## References

1. Berardi, D., Calvanese, D., De Giacomo, G.: Reasoning on UML Class Diagrams. *Artificial Intelligence* 168(1-2), 70–118 (2005)
2. Brown, A.: An introduction to Model Driven Architecture. IBM Technical Report (2004), <http://www.128.ibm.com/developerworks/rational/library/3100.html>
3. Cali, A., Calvanese, D., De Giacomo, G., Lenzerini, M.: Reasoning on UML Class Diagrams in Description Logics. In: Proc. of IJCAR Workshop on Precise Modelling and Deduction for Object-oriented Software Development, PMD 2001 (2001)
4. De Giacomo, G., Lenzerini, M., Rosati, R.: Towards higher-order DL-Lite (preliminary report). In: Proceedings of the International Workshop on Description Logic (DL 2008), Dresden, Germany, May 13-16 (2008)

---

<sup>1</sup> <http://www.trowl.eu>



5. Jekjantuk, N., Gröner, G., Pan, J.Z.: Reasoning in Metamodeling Enabled Ontologies. In: Proceeding of the International workshop on OWL: Experience and Directions, OWL-ED 2009 (2009)
6. Motik, B.: On the properties of metamodeling in owl. *J. Log. Comput.* 17(4), 617–637 (2007)
7. Motik, B., Patel-Schneider, P.F., Parsia, B.: Owl 2 web ontology language: Structural specification and functional-style syntax. World Wide Web Consortium, Working Draft WD-owl2-semantic-20081202 (December 2008)
8. Pan, J.Z., Horrocks, I.: RDFS(FA) and RDF MT: Two Semantics for RDFS. In: Fensel, D., Sycara, K., Mylopoulos, J. (eds.) ISWC 2003. LNCS, vol. 2870, pp. 30–46. Springer, Heidelberg (2003)
9. Pan, J.Z., Horrocks, I., Schreiber, G.: OWL FA: A Metamodeling Extension of OWL DL. In: Proceeding of the International Workshop on OWL: Experience and Directions, OWL-ED 2005 (2005)
10. Patel-Schneider, P.F., Hayes, P., Horrocks, I.: OWL Web Ontology Language Semantics and Abstract Syntax. Technical report, W3C, W3C Recommendation (February 2004)
11. UML. Unified Modeling Language, <http://www.uml.org/>
12. Van Der Straeten, R., Simmonds, J., Mens, T.: Detecting Inconsistencies between UML Models using Description Logic. In: Proceedings of the 2003 International Workshop on Description Logics (DL 2003), Rome, Italy, vol. 81, pp. 260–264 (2003)
13. W3C. Extensible Markup Language (XML) (2001), <http://www.w3.org/XML/>

# Towards Encoding Background Knowledge with Temporal Extent into Neural Networks

Han The Anh and Nuno C. Marques

Centro de Inteligência Artificial (CENTRIA)  
Departamento de Informática, Faculdade de Ciências e Tecnologia,  
Universidade Nova de Lisboa, 2829-516 Caparica, Portugal  
h.anh@fct.unl.pt, nmm@di.fct.unl.pt

**Abstract.** Neuro-symbolic integration merges background knowledge and neural networks to provide a more effective learning system. It uses the Core Method as a means to encode rules. However, this method has several drawbacks in dealing with rules that have temporal extent. First, it demands some interface with the world which buffers the input patterns so they can be represented all at once. This imposes a rigid limit on the duration of patterns and further suggests that all input vectors be the same length. These are troublesome in domains where one would like comparable representations for patterns that are of variable length (e.g. language). Second, it does not allow dynamic insertion of rules conveniently. Finally and also most seriously, it cannot encode rules having preconditions satisfied at non-deterministic time points – an important class of rules. This paper presents novel methods for encoding such rules, thereby improves and extends the power of the state-of-the-art neuro-symbolic integration.

## 1 Introduction

This paper presents new methods for encoding rules with temporal extent into neural networks and explores its usage to improve the neuro-symbolic (NeSy) integration [9].

It has been recently shown that the merging of theory (background knowledge) and data learning (learning from examples) in neural networks provides a more effective learning system [9]. Prior embedding of background knowledge into neural networks leads to faster convergence in many problem domains. The background knowledge is coded in terms of propositional rules, which are then inserted into the neural networks using the Core Method techniques [3,9].

The Core Method was presented for the first time in [3], for computing the least fix-point of normal propositional logic programs in a single hidden layer neural network. Basically, the network is set up such that for each rule of the program, an output unit representing the head of the rule is activated if and only if the input units representing its preconditions are so, simultaneously. The truth value of preconditions of rules must be known from each input pattern. However, in most practical problems, an input pattern determines the value of at most one precondition. In order for all preconditions to be expressed, the input layer must have a big enough size to buffer them. However, this cannot be done if (1) the rules to be encoded are not known at the beginning; or (2) the preconditions are determined at non-deterministic time points. The first issue

prevents the state-of-the-art NeSy integration from allowing dynamic insertion of new rules into the network. This is undesirable because there could be new rules learned during the training, which need to be embedded to improve the network (e.g. in the work on guiding backpropagation method [4]). The second one is more serious. It prevents the NeSy integration from encoding rules having preconditions being determined at non-deterministic time. However, this class of rules is ubiquitous and important, especially in Natural Language Processing (NLP) (e.g. Context Free Grammar (CFG) rules [11]).

In addition, the approach of buffering input patterns so that the patterns can be checked all at once is known as the explicit approach to representing time in neural networks [2]. This approach itself has several drawbacks [2]. One of them is that it demands some interface with the world which buffers the input so it can be represented simultaneously. This imposes a rigid limit on the duration of patterns (since the input layer must provide for the longest possible pattern), and further suggests that all input vectors be the same length. These problems are particularly troublesome in domains such as language, where one would like comparable representations for patterns that are of variable length.

To resolve those drawbacks, we propose a novel approach for encoding rules with preconditions being satisfied at different time points, deterministic as well as non-deterministic, into Elman-based neural networks [2]. The idea is that we provide neural networks with a memory in order to keep track of the network states when older patterns were given—while waiting for the new comming patterns—so that necessary patterns can be gathered and checked all at once.

In the sequel we briefly review the Elman neural network model and the Core Method. Section 3 describes new methods for encoding rules with temporal extent into neural networks. In Section 4, Part-of-Speech tagging – a real world application, is described for illustration. The paper ends with conclusions and future work directions.

## 2 Preliminaries

### 2.1 Elman–Based Recurrent Neural Network

We briefly recall the Elman neural network model – a special kind of recurrent neural networks [12]. In [2], Elman presented a network model to deal with problems of representing time in neural networks, namely of dealing with patterns that have temporal extent. He argued that the explicit approach that associates the serial order of the pattern with the dimensionality of the pattern vector (i.e. the first temporal event is represented by the first element in the pattern vector, the second temporal event is represented by the second position in the pattern vector; and so on) has several drawbacks (see [2] for details). He then described an *implicit* approach to overcome those drawbacks.

Elman’s approach can be briefly depicted as follows. He introduced a three-layer network with the addition of a set of *context units* in the input layer. There are connections from the hidden layer to these context units fixed with a weight of one. At each time step, the input is propagated in a standard feedforward fashion, then a learning rule is applied. The fixed back connections result in the context units always maintaining a copy of the previous values of the hidden units (since they propagate over the connections before the learning rule is applied). Thus the network can maintain a sort of state,

allowing it to perform such tasks as sequence-prediction that is beyond the power of a standard multilayer feedforward networks.

The insight is that Elman equipped the network with a memory in order to (implicitly) represent time by the effect it has on processing. The additional *context units* do not participate in training phase. They are used only to memorize the state of some units, then input to some units (which are not necessarily the ones it memorized for) in the next step. Moreover, they can have a self-connection to enable a long-term memory.

## 2.2 Core Method

The Core Method was introduced in [3], showing how to encode normal programs into neural networks of *binary threshold* units (i.e. units with a step-wise activation function). It has been shown to be useful for encoding knowledge into neural networks [3,4,7], enabling them to achieve faster convergence and better precision results than standard non-symbolic neural networks. In the sequel we briefly recall the method. A full discussion can be found, e.g. in [3,9]. To begin with, let us recall some definitions.

**Definition 1.** *An atom is a propositional variable. A literal is an atom or a negated atom. A normal program  $P$  is a collection of rules of the form  $A \leftarrow L_1, \dots, L_n$ , where  $n \geq 0$ ,  $A$  is an atom and  $L_i, 1 \leq i \leq n$ , are literals.*

**Definition 2.** *A valuation is a mapping from a set of ground atoms to  $\{true, false\}$ . It is extended to literals, rules and programs in the usual way.*

**Definition 3.** *The function  $T_P$  mapping valuations to valuations is defined as follows. Let  $v$  be an valuation and  $A$  an atom.  $T_P(v)(A) = true$  iff  $\exists A \leftarrow L_1, \dots, L_n \in P$  such that  $\bigwedge_{i=1}^n v(L_i) = true$ .*

Let  $P$  be a normal program, and  $n$  and  $m$  be the number of variables and number of rules of  $P$ , respectively. The 3-layer feedforward network constructed as follows computes  $T_P$  operator

1. The input and output layer is a vector of binary threshold units of length  $n$ , representing the  $n$  variables. The thresholds of units occurring in the input and output layers are set to  $.5$  and  $0.5w$ , respectively, where  $w$  is a user-defined constant.
2. For each rule of the form  $A \leftarrow L_1, \dots, L_k, k \geq 0$ , occurring in  $P$  do the following
  - (a) Add a binary threshold unit  $c$  to the hidden layer.
  - (b) Connect  $c$  to the unit representing  $A$  in the output layer with weight  $w$ .
  - (c) For each  $L$  occurring in  $L_1, \dots, L_k$  connect the unit representing  $L$  in input layer to  $c$ . If  $L$  is an atom, set the weight to  $w$ ; otherwise, set the weight to  $-w$ .
  - (d) Set the threshold  $\theta_c$  of  $c$  to  $(l - 0.5)w$ , where  $l$  is the number of positive literals occurring in  $L_1, \dots, L_k$ .

*Core Method for Sigmoid Neural Networks.* The best-known algorithm for training feedforward neural networks is error backpropagation [12], which deals with continuous values. However, as in the Core Method all units of the encoding neural network need to be binary threshold, disabling the algorithm to be applied properly. Fortunately,

e.g. in [5], it has been proven that the Core Method works with sigmoid neural networks (i.e. the network with sigmoid units<sup>1</sup>).

**Proposition 1.** *Any logic program with Boolean inputs can be encoded in a sigmoid neural network.*

Details of the proof can be found in [5]. The idea is that we can make the user-defined constant  $w$  big enough so that the sigmoid function approximates the step-wise function, up to a small enough value. This guarantees that all the sigmoid units of the network behave (activate and prohibit) in the same way as binary threshold units. This Proposition enables us to use backpropagation for training Core Method neural nets.

*Drawbacks of Core Method based Neuro-Symbolic Integration* The network construction implies that the truth value of preconditions of each rule must be known from each input pattern. However, in most practical problems, e.g. Part of Speech (PoS) tagging [1,8], it is not desirable if not achievable because usually each input pattern only determines the value of one precondition (e.g. in PoS tagging the tag of an input word is unique). In order for all preconditions to be determined at once, the input layer must have a big enough size to buffer the input patterns until the value of all preconditions can be determined. This can be done if the following two conditions are satisfied

1. The rules to be encoded are known before constructing the neural network; and
2. For each rule, the time points when its preconditions are determined are known.

The reason for the first condition is obvious. If the rules to be encoded are not known before constructing the network, how can we know the size of the input layer? Although we could define a very big window size, it is very likely to give some overhead to the network performance since there would be more parameters/weights to be learned. Using a bigger size of input patterns leads to worse generalization error, and implies worse performance on testing data [13].

An example where this condition is not satisfied is when new rules are learned on the fly and need to be embedded again into the network (e.g. for faster convergence). It can happen that one of these rules has too many preconditions, more than the size of the buffer, thus impossible to be inserted using the Core Method. This issue actually appeared in the NeSy method for guiding backpropagation learning described in [4]. New rules are learned from wrongly classified patterns (after some training), then being inserted into the network in order to cover those patterns, therefore making the training of backpropagation faster and avoiding local minima.

The reasons for the second condition is that if the time points when the preconditions are determined are not known, first of all, like for the first condition, the size of the input layer is not known before constructing the network. More seriously, even if we gave it very big size, we wouldn't know which units of the input layer determine which preconditions. This issue is very ubiquitous, namely in NLP. For example, we may want to encode the rule saying that the sequence of a determiner, followed by as many adverbs and adjectives as possible, then followed by a noun, forms a noun phrase.

---

<sup>1</sup> A sigmoid unit is the one that has a sigmoid activation function, which is of the form  $f(x) = \frac{1}{1+e^{-x}}$ .

In the next section we present methods to overcome the mentioned drawbacks of the Core Method, namely:

- It is not required to buffer input patterns for dealing with rules having preconditions satisfied at different time points. Hence, new rules can be dynamically inserted into the network.
- It can encode rules with preconditions satisfied at non-deterministic time.

### 3 Rules with Temporal Extent

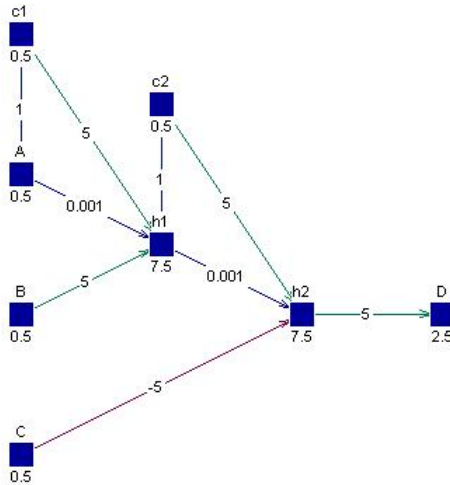
Rules with temporal extent are the ones that have preconditions not being satisfied simultaneously. Each input pattern determines the truth value of at most one precondition. Thus, the value of all of its preconditions can be determined only after a sequence of input patterns. Depending on the property of time points when preconditions are determined, we categorize these rules into three classes and provide an encoding method for each of them. To begin with, let us consider the simplest class of rules that have preconditions satisfied at consecutive time points.

**Definition 4 (Tempo Rule).** *The rule of the form*

$$H \leftarrow_t L_1, L_2, \dots, L_n \quad (n \geq 0)$$

where  $H$  is an atom and  $L_1, \dots, L_n$  are literals, saying that  $H$  is true at time point  $n+T$  ( $T \geq 0$ ) iff  $L_k$  are so at time point  $k+T$ ,  $1 \leq k \leq n$ , is called a tempo rule.

We show how *tempo* rules can be encoded into an Elman-based neural network. Let us consider  $(n+1)$ -layer network  $T$  constructed as follows, where  $w$  is a user-defined number and  $\epsilon$  is a very small number.



**Fig. 1.** Neural Network Encoding Rule:  $D \leftarrow_t A, B, \text{not } C$  where  $w = 5$  and  $\epsilon = 0.001$

1. Input layer has  $n$  units,  $L_1, \dots, L_n$ . Each of them is binary threshold if it is an atom, and identity<sup>2</sup> otherwise. Output layer has a binary threshold unit  $H$ . The threshold of each unit of the input layer is set to  $.5$ , and of the output one is set to  $0.5w$ .
2. Add  $(n - 1)$  one-unit hidden layers  $h_i$  and  $(n - 1)$  context units  $c_i, 1 \leq i \leq n - 1$ .
3. For  $1 \leq i \leq n - 2$ , connect  $L_i$  to  $h_{i-1}$  with weight  $w$  if  $L_i$  is an atom and  $-w$  otherwise. Connect  $L_1$  to  $h_1$  and  $h_i$  to  $h_{i+1}, 1 \leq i \leq n - 2$ , with weight  $\epsilon$ .
4. Connect  $L_1$  to  $c_1, h_i$  to  $c_{i+1}, 1 \leq i \leq n - 2$ , with weight  $1$ . Connect  $c_1$  to  $h_1$  with weight  $w$  if  $L_1$  is an atom and  $-w$  otherwise. If  $n \geq 3$ , connect  $c_i$  to  $h_i, 2 \leq i \leq n - 1$ , with weight  $w$ .
5. Connect  $h_{n-1}$  to the only output unit with weight  $w$ .
6. For  $1 \leq i \leq n - 1$ , set threshold of  $h_i$  to  $1.5w$ .

Note that input of the network is  $n$ -dimensional unit vectors, representing that at each time point the value of only one precondition can be determined.

**Definition 5.** We say a vector is an  $n$ -dimensional unit vector if one of its components is  $1$  or  $-1$ , and all the others are  $0$ . We denote by  $e_i, i = 1, \dots, n$ , the  $n$ -dimensional unit vector component at position  $i$  of which equals  $1$ .

*Example 1.* Encode the tempo rule:  $D \leftarrow_t A, B, \text{not } C$

The network for encoding this rule is depicted in Figure 1, where  $w = 5$  and  $\epsilon = 0.001$ .

**Theorem 1.** Neural network  $P$  is activated (outputs  $1$ ) iff the sequence of  $n$ -dimensional vectors  $a_1, \dots, a_n$  where  $a_i = e_i$  if  $L_i$  is an atom and  $a_i = -e_i$  otherwise, is presented at the input.

*Proof (sketch).* We can prove by induction by  $k$  ( $k \geq 2$ ) that  $h_{k-1}$  is activated if and only if the sequence  $a_1, \dots, a_k$  is presented at the input. In fact, for the base case, it can be easily proved for  $k = 2$ . The induction step can be shown by observing that  $h_k$  is activated at time point  $t$  if and only if  $c_k$  outputs  $1$  at  $t$ , i.e.  $h_{k-1}$  is activated at  $t - 1$ , and at  $t$ ,  $L_{k+1}$  outputs  $1$  if it is an atom and  $-1$  otherwise.  $\square$

For *tempo* rules, the preconditions must be satisfied at consecutive time points. They can be generalized to have preconditions satisfied at arbitrarily known time points.

**Definition 6 (Generalized Tempo Rule).** The rule of the form

$$H \leftarrow_{gt} L_1[t_1], \dots, L_n[t_n] \quad (t_1 < \dots < t_n, n \geq 0)$$

where  $H$  is an atom and  $L_1, \dots, L_n$  are literals, saying that  $H$  is true iff  $L_k$  is true at time point  $t_k + T$  ( $T \geq 0$ ) and not true at any time point  $t$  in between  $t_{k-1} + T$  and  $t_k + T, 1 \leq k \leq n$ , is called a generalized tempo rule.

In order to encode a *generalized tempo* rule we can simply add empty preconditions at time points absent between each two consecutive time points in the rule, then use the same method for *tempo* rules, with one modification that we do not add any input units for the empty preconditions. Below we change Example 1 for illustration.

<sup>2</sup> An identity unit is the one that has an identity activation function.

*Example 2.* Encode the generalized tempo rule:  $D \leftarrow_{gt} A[1], (not\ C)[3]$

This rule states that  $D$  is true if  $A$  is true at time point  $T+1$  and  $(not\ C)$  at  $T+3$  (for any  $T \geq 0$ ). We add an empty precondition at  $T+2$  and obtain *tempo* rule:  $D \leftarrow_t A, Empty, not\ C$ . This rule is encoded with the network in Figure 1 with unit  $B$  being removed and threshold of  $h1$  being changed to 2.5.

**Theorem 2.** *Let  $G$  be the neural network encoding the generalized tempo rule in Definition 6. Then,  $G$  is activated (outputs 1) iff the pattern vectors  $a_1, \dots, a_n$  where  $a_i = e_i$  if  $L_i$  is an atom and  $a_i = -e_i$  otherwise, are presented at the input at time points  $t_1, \dots, t_n$ , respectively. In addition, for  $1 \leq i \leq n$ , there is no input vector  $a_i$  at time point in between  $t_{i-1}$  and  $t_i$ .*

*Proof (sketch).* This theorem can be proved similar to Theorem 1.  $\square$

For *generalized tempo* rules, their preconditions can be satisfied at arbitrarily time points, but the time points must be deterministic. Next, we define and provide a method to encode rules with preconditions satisfied at non-deterministic time.

**Definition 7 (Non-deterministic Tempo Rule).** *The rule of the form*

$$H \leftarrow_{nd} L_1, [[\theta_1]], L_2, \dots, [[\theta_{n-1}]], L_n \quad (n \geq 0)$$

where  $H$  is an atom,  $L_1, \dots, L_n$  are literals and  $\theta_1, \dots, \theta_{n-1}$  are sets of atoms, saying that  $H$  is true iff for all  $k$ ,  $1 \leq k \leq n-1$ ,  $L_{k+1}$  is true after the time  $L_k$  is true, plus that no atom in  $\theta_k$  is true in between, is called a non-deterministic tempo rule.

This rule can be encoded as follows

1. Let  $E$  be the network encoding the tempo rule  $H \leftarrow_t L_1, \dots, L_n$ . Apply to  $E$  the follow steps.
2. For  $1 \leq i \leq n-1$ , add a self-connection to  $c_i$  with weight 1.
3. Let  $S = \cup_{i=1}^{n-1} \theta_i \setminus \{L_1, \dots, L_n\}$  (i.e.  $S$  is the set atoms that appear in some  $\theta_i$ ,  $1 \leq i \leq n-1$ , and does not appear in  $\{L_1, \dots, L_n\}$ ). For each atom in  $S$  add a binary threshold input unit with threshold .5.
4. For  $1 \leq i \leq n-1$ , connect each input unit representing an atom  $A \in \theta_i$  to  $c_i$  with weight -1.

Let  $ND$  be the obtained neural network. It has  $m = n + |S|$  input units. The input vectors of the network are  $m$ -dimensional unit vectors, components of which represent literals in the set  $S \cup \cup_{i=1}^{n-1} \theta_i$ . Let  $pos(L)$  denote the position of the component representing  $L$ ,  $L \in S \cup \cup_{i=1}^{n-1} \theta_i$ .

**Theorem 3.**  *$ND$  is activated (outputs 1) iff the  $m$ -dimensional pattern vectors  $a_{pos(L_1)}, \dots, a_{pos(L_n)}$  where  $a_{L_i} = e_{L_i}$  if  $L_i$  is an atom and  $a_{L_i} = -e_{L_i}$  otherwise, are presented at the input, one after one (not necessary consecutively), such that for  $1 \leq i \leq n$ , there is no input vector  $e_{pos(A)}$ ,  $A \in \theta_i$ , in between the presences of  $a_{pos(L_{i-1})}$  and  $a_{pos(L_i)}$ .*



*Proof (sketch).* We can prove by induction by  $k$  ( $k \geq 2$ ) that  $h_{k-1}$  is activated iff  $a_{pos(L_1)}, \dots, a_{pos(L_k)}$  are presented at the input, one after one, such that for  $1 \leq i \leq k$ , there is no input vector  $e_{pos(L_i)}, L \in \theta_i$ , in between the presences of  $a_{pos(L_{i-1})}$  and  $a_{pos(L_i)}$ . This is done by the following observations. Unit  $h_k$  is activated at some time point  $t$  iff at  $t$ ,  $L_{k+1}$  outputs 1 if it is an atom and -1 otherwise, and  $c_k$  outputs 1. Since according to the network construction,  $c_k$  has input connections from  $h_{k-1}$  (weight 1), from itself (weight 1) and input units representing atoms of  $\theta_{k-1}$  (weights 1), the later holds iff  $h_{k-1}$  is activated at some time point  $t' < t$  and no input unit representing atoms of  $\theta_{k-1}$  outputs 1 in between  $t'$  and  $t$ .  $\square$

Note that for these three rules, the head of a rule is activated at the time the last precondition in its body is activated. When the body of a rule is empty, we call it a *tempo* fact, which is of the form  $A[t]$ . It says  $A$  is true at time point  $t$ .

**Definition 8 (Tempo Logic Program).** A *tempo normal logic program (TNLP for short)* consists of a set of rules of the three forms defined above, and a set of *tempo* facts.

A TNLP is encoded into a neural network by combining the networks that encode each rule of the program using the corresponding method. Next we define a  $T_P$  operator for TNLPs by a program transformation into normal logic programs.

*Program Transformation for TNLP.* Let  $Q$  be a TNLP. Let  $\phi(Q)$  be the NLP obtained from  $Q$  as follows

- For each *tempo* fact  $A[t]$  of  $Q$ , the fact  $A(t)$  in  $\phi(Q)$ .
- For each *tempo* rule  $H \leftarrow_t L_1, L_2, \dots, L_n$  ( $n \geq 1$ ) in  $Q$ , the rule

$$H(T+n) \leftarrow L_1(T+1), L_2(T+2), \dots, L_n(T+n)$$

is in  $\phi(Q)$ .

- For each *generalized tempo* rule  $H \leftarrow_{gt} L_1[t_1], \dots, L_n[t_n]$  ( $n \geq 1$ ) the rule

$$H(t_n + T) \leftarrow L_1(t_1 + T), \dots, L_n(t_n + T)$$

is in  $\phi(Q)$ .

- For each *non-deterministic tempo* rule  $H \leftarrow_{nd} L_1, [[\theta_1]], L_2, \dots, [[\theta_{n-1}]], L_n$  ( $n \geq 1$ ) the following rules are in  $\phi(Q)$  (suppose  $\theta_i = \{Z_{i1}, \dots, Z_{im_i}\}$ ,  $m_i \geq 0$ )

$$H(T_n) \leftarrow L_1(T_1), \text{ not } \theta_1(T_1, T_2), L_2(T_2), \dots, \text{ not } \theta_{n-1}(T_{n-1}, T_n), L_n(T_n), \\ T_1 < T_2 < \dots < T_n.$$

$$\theta_i(T_i, T_{i+1}) \leftarrow Z_{i1}(TZ_{i1}), \dots, Z_{im_i}(TZ_{im_i}), \\ T_i < TZ_{i1} < \dots < TZ_{im_i} < T_{i+1}. \quad (\text{for } 1 \leq i \leq n-1)$$

Note that in the argument of literals, capital letters denote variables. They stand for all possible instantiations with elements of the Herbrand universe.

$T_P$  operator for the TNLP  $Q$  is defined as the usual  $T_P$  operator applied to its NLP counterpart  $\phi(Q)$ . It is easily seen that the following theorem holds

**Theorem 4.** *For each tempo normal logic program  $Q$  there exists a neural network which computes its  $T_P$  operator.*

Similar to the original Core Method, this result provides an approach to designing and implementing a massively parallel computational model for knowledge base—in terms of logic programs—with temporal extent.

## 4 Neuro-symbolic Part-of-Speech Tagging

The methods will be applied to Part-of-Speech (PoS) tagging problem. PoS tagging is an important immediate step in many NLP applications. A PoS tagger is a program that assigns each word in a text a grammatical tag (or category) from a previously defined set – the tagset of that language [14]. The NeSy approach to PoS tagging that conjoins the advantages of background knowledge expressed as rules and the learning power of neural networks [1,8] has been shown to be useful and efficient. However, since this NeSy PoS tagging has recourse to the Core Method for encoding rules, it is confronted with the same drawbacks discussed above. First, it demands a context window to buffer input patterns. Second, it cannot encode rules having preconditions satisfied at non-deterministic time, though they are widespread in PoS tagging domain [1,6,8].

We show how our methods are used to encode rules in PoS tagging into a neural network that does not require a context window for buffering input patterns. We also describe how *non-deterministic tempo* rules in PoS tagging domain—something that was impossible to cope with in current NeSy integration—are encoded within our framework. We show preliminary experimental result to illustrate the usefulness and correctness of the methods.

Let  $\{T_1, \dots, T_n\}$  be a set of tags. We adopt the method in [1] for data generation. Before classifying a word by the network, it is represented by a probability vector. Namely, we assign vector  $(p_w(T_1), \dots, p_w(T_n))$  to word  $w$ , where  $p_w(T_i) = \frac{freq(w, T_i)}{freq(w)}$ . In this equation,  $freq(w, T_i)$  denotes the number of times  $w$  is tagged as  $T_i$  ( $1 \leq i \leq n$ ) and  $freq(w)$  denotes the total number of occurrences of  $w$  in the corpus.

Then, each vector is coupled with an  $n$ -dimensional unit vector (with single 1 for the correct tag) to create a training pattern. The validation data is built from a different subset of the original corpus in the same manner.

The background knowledge is expressed in terms of rules with temporal extent. The most frequent rules extracted from the corpus are shown in Table 1. They are encoded using *tempo* rules. For example, rules 1 and 5 resolve noun-verb ambiguities: the tag (auxiliary verb in rule 1 and pronoun in rule 5) of the first word in the 2-word sequence determines that the most likely tag (noun) should be changed into a verb. In addition, we consider two *non-deterministic tempo* rules for noun phrases (*np*) in Table 2. Rule 1 says, a sequence of words that starts with a determiner and ends with a noun, is a noun phrase, whatever words (not noun) occur in between. The rule 2 says, a sequence of words starting with a wh-word and ends with a noun, plus that there are no verbs or auxiliary verbs in between, is a noun phrase.

**Table 1.** The Six Most Frequent Rules from The Training Corpus

Rule	Example
1) $v \leftarrow_t \text{aux, n}$	would <i>force</i> , would <i>amount</i>
2) $wh \leftarrow_t \text{det, wh}$	<i>a few, a lot, the rest</i>
3) $be \leftarrow_t \text{pront, n}$	it's, there's
4) $wh \leftarrow_t \text{prep, conj}$	of <i>that</i> , during <i>that</i> , like <i>that</i>
5) $v \leftarrow_t \text{pron, n}$	they <i>work</i> , we <i>need</i> , i <i>hope</i>
6) $\text{pront} \leftarrow_t \text{det, aux}$	one <i>might</i> , one <i>can</i>

**Table 2.** Non-deterministic Tempo Rules for Noun Phrases

Rule	Example
1) $\text{np} \leftarrow_{\text{nd}} \text{det, n}$	the <i>cat</i> , the nice <i>cat</i> , the extremely nice <i>cat</i>
2) $\text{np} \leftarrow_{\text{nd}} \text{wh,} \left[ \left[ \{v, \text{aux}\} \right] \right], \text{n}$	this <i>dog</i> , this nice <i>dog</i> , this very nice <i>dog</i> Not correct: which was, which would

For evaluation, we use Susanne corpus [15] (a syntactically annotated version of the Brown corpus – one of the most typical corpora in PoS tagging) and different Elman networks. The networks are deployed and run in JavaNNS software [10].

The original tagset of Susanne corpus is too big for the purposes of this paper: the morphological information is very detailed, and many tags occur only once or twice in the corpus, thus potentially causing problems due to data sparseness. We use a reduced tagset containing only 18 tags, shown in Figure 2.

We embed the rules from Tables 1 and 2 into a neural network using the described methods, and add into it 10 more hidden units for learning other rules. We refer to this network as *CoreNet*. Another network, with the same architecture, is randomly initialized. This network is referred to as *RandNet*.

Both networks are trained on a small dataset (750 patterns). A validation set (45000 patterns) is used to avoid overfitting. Then, a test set (40000 patterns) is used for comparing tagging performance.

Tag	Meaning	Frequency	Tag	Meaning	Frequency
adj	adjective	6022	adv	adverb	4460
aux	auxiliary verb	1055	be	verb to be	2852
conj	conjunction	4124	det	determiner	10829
do	verb to do	294	have	verb to have	856
int	interrogative pronoun	131	n	noun	22493
pto	punctuation	14605	pn	proper name	457
prep	preposition	11050	pron	pronoun	1332
pront	pronoun (3rd person singular)	2544	vt	verb (3rd person singular)	548
v	verb	8476	wh	wh-word	1895

**Fig. 2.** Table of Susanne Reduced Tagset

To evaluate the tagging performance, we tag the test set using the best trained *CoreNet* and *RandNet*. The tagging precision and recall for different tags, overall precision and overall recall are given in Table 3. For a given tag, TP, FP and FN stand for True Positive (correct tag), False Negative (the tagger did not classify the word with the tag, but

**Table 3.** Tagging Precision for Networks Trained with a Small Dataset

		det	n	v	adj	adv	aux	wh	be	np	overall
CoreNet	TP	5477	11938	3449	2918	1610	508	799	1412	219	
	FP	28	163	257	231	268	3	107	8	1	
	FN	10	403	103	158	215	3	891	4	33	
	$Precision = \frac{TP}{TP+FP}$	0.99	0.98	0.93	0.92	0.85	0.99	0.88	0.99	0.88	0.95
	$Recall = \frac{TP}{TP+FN}$	0.98	0.97	0.97	0.95	0.88	0.98	0.47	0.99	0.99	0.95
RandNet	TP	5472	11934	3442	2902	1614	508	0	1412	219	
	FP	33	167	264	247	264	3	906	8	2	
	FN	105	411	97	161	1741	7	0	4	32	
	$Precision = \frac{TP}{TP+FP}$	0.99	0.98	0.92	0.92	0.84	0.98	0	0.99	0.87	0.92
	$Recall = \frac{TP}{TP+FN}$	0.98	0.96	0.97	0.94	0.48	0.98	NaN	0.99	0.99	0.92

it should), False Positive (the tagger classified the word with the tag, but it was wrong), respectively. Note that since all words in the corpus are relevant and tagged (retrieved), we have overall recall equal to overall precision. According to the table, *CoreNet* is, in most cases, better than *RandNet*. Most rules are not so discriminative regarding the training patterns that occur in the corpus. However all the tags with some supporting rule attained some improvement, especially for the *wh* tag (88% vs 0%), that probably has a badly represented pattern in the train set. We should also notice that results for the *np* tag are still preliminary since np-rules seem to need more coverage for other noun phrases and we have a (comparatively) small number of labeled samples.

The better performance of *CoreNet* compared with *RandNet* is because of the effect of the encoded rules. They enable to quickly classify and learn the patterns that they encode. Also, in our experiments, we observed that the network initialized with rules attains the best state much faster than the randomly initialized one. This, to some extent, suggests a similarity to the learning process of our biological neural net. Being equipped with some background knowledge it can learn faster than starting from nothing.

## 5 Conclusions and Future Work

The state-of-the-art NeSy integration that has recourse to the Core Method for encoding rules has several limitations in dealing with rules having temporal extent. First, it demands some interface with the world which buffers the input so it can be represented all at once. This imposes a rigid limit on the duration of patterns and requires all input vectors be the same length. These are troublesome in domains such as language, where one would like comparable representations for patterns that are of variable length. Second, it does not handle well the dynamic insertion of rules into the network; though this ability is much desirable and important, for there might be new rules learned during the training that their insertion could enhance the network performance. And most seriously, it cannot encode *non-deterministic tempo* rules while this class of rules is ubiquitous and very important, e.g. CFG grammar rules in NLP.

To resolve all those drawbacks, we have presented methods for encoding different classes of rules with temporal extent into Elman-based neural networks: (1) Tempo rule – the simplest class where rules have preconditions satisfied at consecutive time points; (2) Generalized tempo rule – this class generalizes the first one to allow preconditions

satisfied at arbitrary, but known, time points; and (3) Non-deterministic tempo rule – rules in this class can have preconditions satisfied at non-deterministic time points.

In short, being the first contribution, with these methods, we have provided the state-of-the-art NeSy integration with a genuinely proper way to handle rules of the first and second classes, and extended its power to deal with rules in the third one. In addition, being another contribution, the methods themselves are useful to encode an important kind of rules – rules with temporal extent, into neural networks. However, its usefulness for other issues, e.g. parallel processing of knowledge bases with temporal extent, is remained for future exploration.

For illustrating the usefulness of our methods, we have shown how they can be applied to develop a NeSy PoS tagger that does not require a context window of input patterns. As shown in the experiments, the tagging performance is good w.r.t. the networks initialized with rules, even with a small training dataset. It also suggests that the Elman network architecture is more appropriate to learn rules with temporal extent.

We have also provided some *non-deterministic tempo* rules of PoS tagging and shown how they are embedded into neural networks. This shows that our methods can handle well rules that can be related to CFG grammars. In future work, this relation should be better studied.

We believe that, with the methods presented in this paper, we have made an important step towards encoding grammar rules. As we, human beings, can do natural language processing effectively after having learned grammar rules at school, it is likely that embedding those rules as background knowledge into neural networks is an appropriate way to make them more efficient in NLP.

## References

1. Marques, N.C., Bader, S., Rocio, V., Hölldobler, S.: Neuro-Symbolic Word Tagging. In: Workshop on Text Mining and Applications, Portuguese Conf. on Artificial Intelligence. IEEE, Los Alamitos (2007)
2. Elman, J.L.: Finding Structure in Time. *Cognitive Science* 14, 179–211 (1990)
3. Hölldobler, S., Kalinke, Y.: Towards a massively parallel computational model for logic programming. In: Workshop on Combining Symbolic and Connectionist Processing. ECAI, pp. 68–77 (1994)
4. Bader, S., Hölldobler, S., Marques, N.C.: Guiding backprop by inserting rules. In: Procs. 4th Intl. Workshop on Neural-Symbolic Learning and Reasoning (2008)
5. Marques, N.C.: An Extension of the Core Method for Continuous Values: Learning with Probabilities. In: Procs. 14th Portuguese Conf. on Artificial Intelligence, pp. 319–328 (2009)
6. Marques, N.C., Lopes, J.G.: Using Neural Nets for Portuguese Part-of-Speech Tagging. In: Procs. the 5th Intl Conf. on Cognitive Science of Natural Language Processing, Ireland (1996)
7. Bader, S., Hitzler, P., Hölldobler, S.: Connectionist model generation: A first-order approach. *Neurocomputing* 51, 2420–2432 (2008)
8. Marques, N.C., Lopes, G.P.: Neural networks, part-of-speech tagging and lexicons. In: Hoffmann, F., Adams, N., Fisher, D., Guimarães, G., Hand, D.J. (eds.) IDA 2001. LNCS, vol. 2189, pp. 63–72. Springer, Heidelberg (2001)
9. d’Avila Garcez, A.S., Broda, K.B., Gabbay, D.M.: Neural- Symbolic Learning Systems Foundations and Applications. In: Perspectives in Neural Computing, Springer, Berlin (2002)

10. Zell, A.: SNNS, stuttgart neural network simulator, user manual, version 2.1. Technical report, Stuttgart (1992)
11. Pereira, F.C.N., Shieber, S.M.: A Prolog and natural-language analysis. CSLI Lecture Notes, vol. 10 (1987)
12. Mitchell, T.M.: Machine Learning. McGraw-Hill, New York (March 1997)
13. Haykin, S.: Neural networks: a comprehensive foundation. Prentice Hall, Englewood Cliffs (1999)
14. van Halteren, H.: Syntactic Wordclass Tagging. Kluwer Academic Publishers, Dordrecht (1999)
15. Sampson, G.: English for the Computer: The SUSANNE Corpus and Analytic Scheme. Oxford University Press, Oxford (1995)

# A Fuzzy Description Logic with Automatic Object Membership Measurement

Yi Cai<sup>1</sup> and Ho-Fung Leung<sup>2</sup>

<sup>1</sup> Department of Computer Science  
The City University of Hong Kong, Hong Kong, China  
yicai3@cityu.edu.hk

<sup>2</sup> Department of Computer Science and Engineering  
The Chinese University of Hong Kong, Hong Kong, China  
lhf@cuhk.edu.hk

**Abstract.** In this paper, we propose a fuzzy description logic named  $f_{om}\text{-}\mathcal{DL}$  by combining the classical view in cognitive psychology and fuzzy set theory. A formal mechanism used to determine object memberships automatically in concepts is also proposed, which is lacked in previous work fuzzy description logics. In this mechanism, object membership is based on the defining properties of concept definition and properties in object description. Moreover, while previous works cannot express the qualitative measurements of an object possessing a property, we introduce two kinds of properties named N-property and L-property, which are quantitative measurements and qualitative measurements of an object possessing a property respectively. The subsumption and implication of concepts and properties are also explored in our work. We believe that it is useful to the Semantic Web community for reasoning the fuzzy membership of objects for concepts in fuzzy ontologies.

## 1 Introduction

Recently, *Description Logics* (DLs) have become popular due to their application on the Semantic Web [1]. While ontologies play an important role in knowledge representation and provide a way to describe and structure the information on the web, DLs are essentially the theoretical counterpart of the *Web Ontology language OWL DL* which is the main language to specify ontologies.

Typically, Description Logics are limited to dealing with crisp concepts [2]. However, there are many vague concepts in reality. These vague concepts have no clear boundaries. For examples, ‘hot water’, ‘red car’ and so on. For these vague concepts, there is not a clear and precise boundary. Description logics and the ontologies based on DLs cannot handle these fuzzy concepts. To extend the representation ability of description logics to handle fuzzy knowledge, some fuzzy description logics have been proposed based on fuzzy logic, such as the fuzzy extension of  $\mathcal{ALC}$  [2], the fuzzy  $SHOIN(\mathcal{D})$  [1] and the fuzzy  $SHIN$  [3]. These fuzzy description logics can handle fuzzy concepts and possess different express power. However, object memberships are to be provided by users manually or obtained by fuzzy functions defined by users in these fuzzy description logics, which are tedious, time-consuming to obtain, and their

effectiveness depends on the user's knowledge on the domain. Psychologists find that there is a more reasonable and natural way to obtain object membership, in which object membership depends on how well the properties of an object satisfy the properties in the definition of a concept.

While concepts, objects and properties are building blocks of ontologies, to our best knowledge, as the theoretical counterpart of fuzzy ontologies, current fuzzy description logics lack a formal expression of properties and a formal mechanism to determine memberships of objects in concepts based on the concept definition and object description. Thus, machines cannot obtain object membership automatically. While properties are generally used in describing concepts and objects in ontologies and people's daily life, it is desirable to have a mechanism to measure object membership in fuzzy description logic based on concept definitions and object descriptions.

In this paper, to overcome the limitations of previous fuzzy description logics, we propose a novel fuzzy description logic named  $f_{om}\text{-}\mathcal{DL}$ , based on theories in cognitive psychology [4] [5] and fuzzy set theory [6]. We provide a formal mechanism to determine object membership in concepts, which is lacked in current fuzzy description logics. The main feature of this mechanism is that object membership is measured by the defining properties of concepts and properties of objects, which is in line with the classical view in cognitive psychology [4] [5]. Besides, two kinds of properties named N-property and L-property are introduced. They are quantitative measurements and qualitative measurements of an object possessing a property, and previous works cannot express and represent the qualitative measurements of an object possessing a property. We also present theories and implications about concept subsumption and property subsumption. It is useful to the Semantic Web community for reasoning the fuzzy membership of objects for concepts in fuzzy ontologies based  $f_{om}\text{-}\mathcal{DL}$ .

The structure of this paper is as follows. Section 2 introduces the background and related work. We introduce the proposed fuzzy description logic  $f_{om}\text{-}\mathcal{DL}$  in section 3. In section 4, we propose a formal mechanism to determine the object memberships in concepts based on the defining properties of concepts and properties in object descriptions. We discuss some interesting features of  $f_{om}\text{-}\mathcal{DL}$  in section 5. Section 6 concludes the paper.

## 2 Background and Related Work

### 2.1 Fuzzy Description Logics and Fuzzy Ontologies

Currently, DLs [7] can only handle crisp concepts but cannot deal with fuzzy concepts. Several fuzzy DLs are proposed to handle fuzzy concepts by combining fuzzy set theory [6] and description logics. For examples, Straccia proposes a fuzzy  $\mathcal{ALC}$  in [2] and a fuzzy  $\mathcal{SHOIN}(\mathcal{D})$  in [1]. Stoilos et al. present a fuzzy  $\mathcal{SHIN}$  in [3]. These fuzzy DLs vary in expressive power, complexity and reasoning capabilities. However, object memberships are given by users manually or they are obtained by fuzzy functions defined by users. Some fuzzy ontologies are constructed based on fuzzy DLs or fuzzy logic [8] [9]. Cai and Leung propose a formal ontology model with property hierarchy and object membership [10], and the work presented in this paper can be seen as a theoretical counterpart for that model.



## 2.2 Classical View in Cognitive Psychology

In cognitive psychology, how concepts are represented in the human memory is an important concern. It is generally accepted that concepts are characterized by properties [11]. One important model of concept representation based on properties is the classical view. The classical view [4] [5] of concepts posits that each concept is defined by a set of properties which are individually necessary and collectively sufficient. Properties are atomic units which are the basic building blocks of concepts. Concepts are organized in a hierarchy and the defining properties of a more specific concept includes all the defining properties of its super-concepts.

## 3 A Description Logic with Automatic Object Membership Measurement

In this section, we present the syntax and semantics of the proposed fuzzy description logic named  $f_{om}\text{-}\mathcal{DL}$ .

### 3.1 Syntax

In  $f_{om}\text{-}\mathcal{DL}$ , we have an alphabet of distinct concepts (**C**), roles (**R**), objects (**I**) and properties (**P**). We adopt the unique name assumption in  $f_{om}\text{-}\mathcal{DL}$ . The syntax of  $f_{om}\text{-}\mathcal{DL}$  is as follows.

**Role.** Each role name  $RN$  is a fuzzy role in  $f_{om}\text{-}\mathcal{DL}$ . A valid role  $R$  is defined by the abstract syntax:  $R := RN | R^-$ . The inverse relation of roles is symmetric, and to avoid considering roles such as  $R^-$ , we defined a function  $\text{Inv}$ , which returns the inverse of a role, more precisely  $\text{Inv}(R) := RN^-$  if  $R = RN$ , and  $\text{Inv}(R) = RN$  if  $R = RN^-$ . Roles are organized in a hierarchy.

**Concept.** Each concept name  $CN \in \mathbf{C}$  is a fuzzy concept in  $f_{om}\text{-}\mathcal{DL}$ . We denote a primitive concept by  $A$ , then concepts  $C$  and  $D$  are formed out as follows.

$$C, D \longrightarrow \top | \perp | A | C \sqcap D | C \sqcup D | \neg C | \forall R.C | \exists R.C | \geq_n R.C | \leq_n R.C | \$R.C | \forall R_1, \dots, R_n.C | \exists R_1, \dots, R_n.C$$

**Object.** Each object name  $IN \in \mathbf{I}$  is an object in  $f_{om}\text{-}\mathcal{DL}$ .

**Property.** Each property name  $PN \in \mathbf{P}$  is a property in  $f_{om}\text{-}\mathcal{DL}$ . A property is defined by the abstract syntax:  $P := R.C$ .  $R$  is a role which is a binary relation between objects and  $C$  is the range concept of  $R$ .

There are two kinds of properties in our fuzzy description logic. One is named **N-Property** while another one is **L-property**. An **N-property** is defined by the following syntax:

$$NP := \exists P | \forall P | \geq_n P | \leq_n P$$

where  $P$  is a property. An **L-property** is defined by the following syntax:  $LP := \$P$  where  $P$  is a property. N-properties are the quantitative measurement of the degrees of objects possessing properties and L-properties are the qualitative measurement of the degrees of objects possessing properties. A **primitive property** is a property in the form of  $R.C$  where  $R$  is a primitive role and  $C$  is a concept consists of only one instance.

According to classical view in cognitive psychology, a concept can be defined by properties. Moreover, the membership of an object in a concept can be measured based on the defining properties of the concept. Borrowing this idea, we consider that there are a **definition** for each concept and a **description** for each object, which are not taken into consideration in previously proposed fuzzy description logics.

**Definition 1.** *The syntax of the **definition** of a concept is as the following form:*

$$\widehat{C} = \vec{S}_{C,1} \sqcup \vec{S}_{C,2} \sqcup \dots \sqcup \vec{S}_{C,m}, 1 \leq i \leq m$$

and

$$\vec{S}_{C,i} = p_{i,1}(r_{i,1}) \sqcap p_{i,2}(r_{i,2}) \sqcap \dots \sqcap p_{i,n_i}(r_{i,n_i}), 1 \leq j \leq n_i$$

where  $n_i$  is the number of properties in  $\vec{S}_{C,i}$ ,  $p_{i,j}$  is a  $N$ -property or a  $L$ -property,  $\vec{S}_{C,i}$  is named a characteristic vector of a concept and  $r_{i,j}$  is a real value in the range of  $[0, 1]$ .

*Example 1.* Suppose an online-shop selects the top 100 special customers to give them some discount. The concept ‘special-customer’ denoted by  $SC$  is a fuzzy concept. One kind of special customers requires that a customer must have bought at least five items (goods) belonging to ‘expensive item’ and the average degree of all items that the customer has bought belonging to ‘expensive item’ is high (and the higher the better). Another kind of special customers includes customers who must have bought at least one hundred items (not necessary expensive items) and at least one item that the customer has bought must belong to ‘expensive item’.  $SC$  can be defined as follows.<sup>1</sup>

$$\widehat{SC} = (\geq_5 \text{ buy.expensiveItem}(1) \sqcap \$\text{buy.expensiveItem}(0.6)) \sqcup (\geq_{100} \text{ buy.Item}(1) \sqcap \geq_1 \text{ buy.expensiveItem}(0.5))$$

For any two concepts, their characteristic vectors (i.e., defining properties in some characteristic vectors) are generally different. To compare the definitions of two concepts  $C$  and  $D$ , we need to align the defining properties of concept  $C$  to that of concept  $D$ . I.e.,  $\widehat{D} = \vec{S}_{C,1}^D \sqcup \vec{S}_{C,2}^D \sqcup \dots \sqcup \vec{S}_{C,m}^D$  where  $\vec{S}_{C,i}^D$  has the same properties as in  $\vec{S}_{C,i}$ . We name each  $\vec{S}_{C,i}^D$  as **aligned characteristic vector** of concept  $D$  for  $C$ .

**Definition 2.** *The syntax of the **description** of an object is as the following form:*

$$\vec{O}_a \cong \exists p_{a,1}(t_{a,1}) \sqcap \exists p_{a,2}(t_{a,2}) \sqcap \dots \sqcap \exists p_{a,n}(t_{a,n}), 1 \leq i \leq n$$

where  $p_{a,i}$  is a primitive property a possessing,  $t_{a,i}$  is the degree to which a possesses property  $p_{a,i}$ . For the reason that all properties in the object property vector are primitive properties, thus  $\forall i, t_{a,i} = 1$ . We consider that an object in an ontology is represented by a set of fuzzy primitive properties named object property vector. The relation among the fuzzy primitive properties in the object property vector is conjunction.

Thus an object  $a$  possesses a set of fuzzy instance properties. In other words, an object can be described by a set of primitive property.

<sup>1</sup> This is used to illustrate the syntax of the concept definition only.

*Example 2.* We assume a customer  $O_1$  has a customer id ‘20071202’ and has bought two items ‘Furniture00002’ and ‘Eproduct00307’.  $O_1$  is described as:

$$\vec{O}_1 \cong \exists hasId.2001202(1) \sqcap \exists buy.Furniture00002(1) \sqcap \exists buy.Eproduct00307(1)$$

**Fuzzy TBox.** A fuzzy *TBox*  $\mathcal{T}$  consists of a finite set of fuzzy concept definition and fuzzy concept inclusion axioms.

**Fuzzy ABox.** A fuzzy *ABox*  $\mathcal{A}$  consists of a finite set of fuzzy (primitive) concept and fuzzy (primitive) role assertion axioms. As for the crisp case,  $\mathcal{A}$  may also contain a finite set of object (in)equality axioms  $a \approx b$  and  $a \neq b$  respectively.

**Fuzzy RBox.** A fuzzy *RBox*  $\mathcal{R}$  consists of a finite set of transitivity axioms  $\text{trans}(\mathbf{R})$  and role inclusion axioms. A *primitive role* is a role that it does not subsume any other roles. In other words, there is no a sub-role for a primitive role.

**Knowledge Base.** A knowledge base  $\mathcal{K} = \langle \mathcal{T}, \mathcal{R}, \mathcal{A} \rangle$  consists of a fuzzy TBox  $\mathcal{T}$ , a fuzzy RBox  $\mathcal{R}$  and a fuzzy ABox  $\mathcal{A}$ .

### 3.2 Semantics

The semantics of the proposed fuzzy  $\mathcal{DL}$  is provided by a fuzzy interpretation which is a pair  $\mathcal{I} = \langle \Delta^{\mathcal{I}}, \cdot^{\mathcal{I}} \rangle$  where the domain  $\Delta^{\mathcal{I}}$  is a non-empty set of objects and  $\cdot^{\mathcal{I}}$  is a fuzzy interpretation function, which maps

- an object name  $\mathbf{a}$  to elements of  $\mathbf{a}^{\mathcal{I}} \in \Delta^{\mathcal{I}}$ ;
- a concept name  $\mathbf{C}$  to a membership function  $\mathbf{C}^{\mathcal{I}} : \Delta^{\mathcal{I}} \rightarrow [0, 1]$ , and we consider the object membership of an object  $a_i$  in a concept  $C$  is denoted by  $\mu_C(a_i)$  and  $\mu_C(a_i) = \mathbf{C}^{\mathcal{I}}(a_i)$ . Thus, a concept  $C$  is considered as a fuzzy set of objects  $C = \{a_1^{u_1}, a_2^{u_2}, \dots, a_n^{u_n}\}$ , where  $a_i$  is an object in  $\Delta^{\mathcal{I}}$  and  $u_i$  is the membership of  $a_i$  in concept  $C$ , i.e.,  $\mathbf{C}^{\mathcal{I}}(a_i)$ ;
- a role name  $\mathbf{R}$  to a membership function  $\mathbf{R}^{\mathcal{I}} : \Delta^{\mathcal{I}} \times \Delta^{\mathcal{I}} \rightarrow [0, 1]$ . A role  $R$  is actually considered as a fuzzy set of object pairs  $R = \{ \langle a_1, b_1 \rangle^{w_1}, \langle a_2, b_2 \rangle^{w_2}, \dots, \langle a_n, b_n \rangle^{w_n} \}$ , where  $\langle a_i, b_i \rangle$  is a role instance (i.e., a pair of objects) and  $w_i$  is the membership of the role instance  $\langle a_i, b_i \rangle$  in  $R$ , i.e.,  $\mathbf{R}^{\mathcal{I}}(a_i, b_i)$ ;
- a property name  $\mathbf{P}$  to a membership function  $\mathbf{P}^{\mathcal{I}} : \mathbf{R}^{\mathcal{I}} \times \mathbf{C}^{\mathcal{I}} \rightarrow [0, 1]$ . For a property, it is interpreted as a fuzzy set of pairs roles and concepts, i.e.,  $P = \{ \langle \langle a_i, b_i \rangle, b_i \rangle^{v_i} \mid \langle a_i, b_i \rangle^{w_i} \in R, b_i^{u_i} \in C \}$ . If an object  $a_i$  has a fuzzy role  $\langle a_i, b_i \rangle$  with object  $b_i$ ,  $R^{\mathcal{I}}(a_i, b_i) = w_i > 0$  and  $C^{\mathcal{I}}(b_i) = u_i > 0$ , then we say  $a_i$  possesses a *property member*  $\langle \langle a_i, b_i \rangle, b_i \rangle$  of property  $P = R.C$  to a degree  $P^{\mathcal{I}}(\langle a_i, b_i \rangle, b_i) = v_i$  where  $P^{\mathcal{I}}(\langle a_i, b_i \rangle, b_i) = \min(R^{\mathcal{I}}(a_i, b_i), C^{\mathcal{I}}(b_i))$ .

For the semantics of the definition of a concept,  $r_{i,j}$  is the minimum requirement of the corresponding property  $p_{i,j}$  in the concept  $C$ . In other words, we consider that  $r_{i,j}$  is an  $\alpha$ -cut of the fuzzy function of property  $p_{i,j}$ . If an object possesses  $p_{i,j}$  to a degree greater than or equal to  $r_{i,j}$ , then we consider that the object satisfies the minimum requirement of property  $p_{i,j}$  of concept  $C$ . For the description of an object, it describes what primitive properties the object possesses.

The complete set of semantics is depicted in table 1. We briefly comment some points for the limitation of space. As mentioned above, there are two kinds of properties which are N-property and L-property.

- An N-property is the quantitative measure for an object  $a$  possessing a property  $p_x$ . It is a number restriction on property members of  $p_x$  that object  $a$  possesses. An N-property is interpreted as a property with a quantifier. There are a set of quantifiers for modelling number restrictions on properties. We present four quantifiers used frequently here, which are  $\exists, \forall, \geq_n, \leq_n$  and  $n$  is an integer. For  $\exists R.C$  (i.e.,  $\exists P$  while  $P = R.C$ ), it is interested as  $\exists R.C^{\mathcal{I}}(a) = \sup_{b \in \Delta^{\mathcal{I}}} \{ \min(R^{\mathcal{I}}(a, b), C^{\mathcal{I}}(b)) \}$ .  $\exists$  is considered as a disjunction over the elements of the domain and thus we use  $\min$  function which is widely adopted for disjunction in fuzzy logic. Similarly,  $\forall R.C^{\mathcal{I}}(a)$  is interpreted as  $\inf_{b \in \Delta^{\mathcal{I}}} \{ \max(1 - R^{\mathcal{I}}(a, b), C^{\mathcal{I}}(b)) \}$  in which  $\forall$  is considered as a conjunction over the elements of the domain, and thus we use the  $\max$  function which is widely adopted for conjunction in fuzzy logic. Furthermore,  $>_n P = \geq_{n+1} P$ ,  $\leq_n P = \neg(>_n P)$  and  $<_n P = \neg(\geq_n P)$ .

*Example 3.* If a customer  $O_c$  has bought a set of items (e.g., ‘Eproduct00307’, ‘Book07005’ and so on). We can use the fuzzy functions defined in table 1 to calculate the degree of  $O_c$  possessing these N-properties. For instance, we can obtain that  $O_c$  possesses the property ‘ $\exists$ buy.Item’ to a degree 1. It means that  $O_c$  definitely has bought at least one item.

- An L-property is the qualitative measure of an object  $a$  possessing a property  $P = R.C$  is a qualitative aggregation on the set of property members of  $P$  which object  $a$  possesses.  $\$$  is interpreted as a qualification aggregation on all property members, and we name it a *qualifier*. There are several aggregation functions to aggregate all the property members [12] [13]. One of the aggregation used frequently for qualitative measure is an average function for membership degrees of property members which objects possess in  $P$  as follows.

$$\$R.C^{\mathcal{I}}(a) = \frac{\sum_{i=1}^n v_i^a}{n} \quad (1)$$

where  $v_i^a$  is the membership degree of property member  $p_i$  of  $P$  object  $a$  possessing.

*Example 4.* Suppose a customer  $O_c$  has bought two items ‘Eproduct00307’ and ‘Furniture00002’ only. Both ‘Eproduct00307’ and ‘Furniture00002’ belong to ‘expensiveItem’ to a degree 1. Then we can obtain that  $O_c$  possesses ‘ $\$$ buy.expensive-Item’ to a degree 1. It means that  $O_c$  definitely has bought expensive items.

### 3.3 Subsumption and Implication

In our fuzzy description logic, concepts, roles and properties are organized in a hierarchy and we define concept subsumption, role subsumption and property subsumption as follows.

**Definition 3.** For two concepts  $X$  and  $Y$ ,  $X = \{a_1^{u_{X,1}}, a_2^{u_{X,2}}, \dots, a_n^{u_{X,n}}\}$  and  $Y = \{a_1^{u_{Y,1}}, a_2^{u_{Y,2}}, \dots, a_n^{u_{Y,n}}\}$ ,  $a_i$  is an object,  $u_{X,i}$  is the membership degree of  $a_i$  in fuzzy concept  $X$  and  $u_{Y,i}$  is the membership degree of  $a_i$  in fuzzy concept  $Y$ . If  $\forall a_i^{u_{X,i}} \in X, a_i^{u_{Y,i}} \in Y, u_{Y,i} \geq u_{X,i}$  then  $X$  is **subsumed by**  $Y$  (or  $Y$  **subsumes**  $X$ ) which is denoted as  $X \subseteq Y$ .

**Table 1.** Semantics of the proposed fuzzy  $DL$ 

$\top^{\mathcal{I}}$	=	1
$\perp^{\mathcal{I}}$	=	0
$(\neg C)^{\mathcal{I}}(a)$	=	$1 - C^{\mathcal{I}}(a)$
$(C \sqcup D)^{\mathcal{I}}(a)$	=	$\max(C^{\mathcal{I}}(a), D^{\mathcal{I}}(a))$
$(C \sqcap D)^{\mathcal{I}}(a)$	=	$\min(C^{\mathcal{I}}(a), D^{\mathcal{I}}(a))$
$(\forall R.C)^{\mathcal{I}}(a)$	=	$\inf_{b \in \Delta^{\mathcal{I}}} \{ \max(1 - R^{\mathcal{I}}(a, b), C^{\mathcal{I}}(b)) \}$
$(\exists R.C)^{\mathcal{I}}(a)$	=	$\sup_{b \in \Delta^{\mathcal{I}}} \{ \min(R^{\mathcal{I}}(a, b), C^{\mathcal{I}}(b)) \}$
$(\geq_n R.C)^{\mathcal{I}}(a)$	=	$\sup_{b_1, \dots, b_n \in \Delta^{\mathcal{I}}} \min_{i=1}^n (R^{\mathcal{I}}(a, b_i), C^{\mathcal{I}}(b_i))$
$(\leq_n R.C)^{\mathcal{I}}(a)$	=	$\neg(\geq_{n+1} R.C)^{\mathcal{I}}(a)$
$(R^-)^{\mathcal{I}}(b, a)$	=	$R^{\mathcal{I}}(a, b)$
$(\$R.C)^{\mathcal{I}}(a)$	=	$\text{aggr}_{1 \leq i \leq n} (P^{\mathcal{I}}(\langle a, b_i \rangle, b_i))$
$(\forall R_1, \dots, R_n.C)^{\mathcal{I}}(a)$	=	$\inf_{b \in \Delta^{\mathcal{I}}} \{ \max(\min_{i=1}^n (1 - R_i^{\mathcal{I}}(a, b)), C^{\mathcal{I}}(b)) \}$
$(\exists R_1, \dots, R_n.C)^{\mathcal{I}}(a)$	=	$\sup_{b \in \Delta^{\mathcal{I}}} \{ \min(\max_{i=1}^n (R_i^{\mathcal{I}}(a, b)), C^{\mathcal{I}}(b)) \}$

**Definition 4.** For two fuzzy roles  $S$  and  $Q$ ,  $S = \{ \langle a_1, b_1 \rangle^{w_{S,1}}, \langle a_2, b_2 \rangle^{w_{S,2}}, \dots, \langle a_n, b_n \rangle^{w_{S,n}} \}$  and  $Q = \{ \langle c_1, d_1 \rangle^{w_{Q,1}}, \langle c_2, d_2 \rangle^{w_{Q,2}}, \dots, \langle c_n, d_n \rangle^{w_{Q,n}} \}$ , if  $\forall \langle a_i, b_i \rangle^{w_{S,i}} \in S, \langle a_i, b_i \rangle^{w_{Q,i}} \in Q, w_{Q,i} \geq w_{S,i}$  then we say  $S$  is **subsumed by**  $Q$  (or  $Q$  **subsumes**  $S$ ) denoted as  $S \subseteq Q$ .  $w_{S,i}$  is the degree of strength of  $\langle a_i, b_i \rangle$  in fuzzy role  $S$  and  $w_{Q,i}$  is the degree of strength of  $\langle a_i, b_i \rangle$  in fuzzy role  $Q$ .

**Definition 5.** For two fuzzy properties  $P_1$  and  $P_2$ ,

$$P_1 = \{ \langle \langle a, c \rangle, c \rangle^{v_{1,i}} \mid \langle a, c \rangle^{w_{1,i}} \in S, c^{u_{1,i}} \in C \}$$

and

$$P_2 = \{ \langle \langle a, c \rangle, c \rangle^{v_{2,i}} \mid \langle a, c \rangle^{w_{2,i}} \in Q, c^{u_{2,i}} \in D \}$$

, if  $\forall \langle \langle a, c \rangle, c \rangle^{v_{1,i}} \in P_1, \langle \langle a, c \rangle, c \rangle^{v_{2,i}} \in P_2, v_{1,i} \leq v_{2,i}$ , then we say  $P_1$  is **subsumed by**  $P_2$  (or  $P_2$  **subsumes**  $P_1$ ) denoted by  $P_1 \subseteq P_2$ .

For the reason that concepts are defined by properties and objects are described by properties also. What is more, in our fuzzy description logic, object membership depends on the match of defining properties of a concept definition and properties of an object description. It is necessary to reason about the concept subsumption and property subsumption. We can obtain two theorems based on semantics axioms and definitions introduced above for concept subsumption and property subsumption.

**Theorem 1.** For two properties  $P_1$  and  $P_2$ , if  $P_1 = S.C, P_2 = Q.D, S \subseteq Q$ , and  $C \subseteq D$ , then  $P_1 \subseteq P_2$ .

*Proof.* For the reason that  $P_1 = \{ \langle \langle a, c \rangle, c \rangle^{v_{1,i}} \mid \langle a, c \rangle^{w_{1,i}} \in S, c^{u_{1,i}} \in C \}$  and  $P_2 = \{ \langle \langle a, c \rangle, c \rangle^{v_{2,i}} \mid \langle a, c \rangle^{w_{2,i}} \in Q, c^{u_{2,i}} \in D \}$ , for each object  $c_i$  in the universe,  $u_{1,i} \leq u_{2,i}$  because  $C \subseteq D$  according to definition 3 about fuzzy concept subsumption. For each object pair  $\langle a, c \rangle, \langle a, c \rangle^{w_{1,i}} \in S$  and  $\langle a, c \rangle^{w_{2,i}} \in Q$ ,  $w_{1,i} \leq w_{2,i}$  because  $S \subseteq Q$  according to definition 4 about fuzzy role subsumption. For each object  $\langle a_i, c_i \rangle, c_i$ ,

$$P_1^{\mathcal{I}}(\langle a_i, c_i \rangle, c_i) = \min(S^{\mathcal{I}}(a_i, c_i), C^{\mathcal{I}}(c_i)) = \min(w_{1,i}, u_{1,i})$$

and

$$P_2^{\mathcal{I}}(\langle a_i, c_i \rangle, c_i) = \min(Q^{\mathcal{I}}(a_i, c_i), D^{\mathcal{I}}(c_i)) = \min(w_{2,i}, u_{2,i}).$$

According to the result that  $w_{1,i} \leq w_{2,i}$  and  $u_{1,i} \leq u_{2,i}$ , thus,  $P_1 \subseteq P_2$  according to the definition 5 about fuzzy property subsumption.

**Theorem 2.** For two concepts  $C$  and  $D$ ,  $\forall \vec{S}_{C,i} \in \widehat{C}$  and  $\vec{S}_{C,i}^D$  is the corresponding aligned characteristic vector of  $D$  for  $C$ , if for each defining property  $x \in \vec{S}_{C,x}$ ,  $r_{i,x}^C$  is the minimum requirement of property  $x$  in  $\vec{S}_{C,i}$  and  $r_{i,x}^D$  is the minimum requirement of property  $x$  in  $\vec{S}_{D,i}$ ,  $r_{i,x}^C \leq r_{i,x}^D$ , then  $D \subseteq C$ .

*Proof.* In each  $\vec{S}_{C,i} \in \widehat{C}$  and its corresponding aligned characteristic vector  $\vec{S}_{C,i}^D$  of  $D$  for  $C$ , for each property  $x$ ,  $r_{i,x}^C \leq r_{i,x}^D$ , it means that the characteristic  $\vec{S}_{C,i}$  has a less restrict requirement on each property  $x$  than  $\vec{S}_{C,i}^D$ , so objects will possess each property to a higher degree in  $\vec{S}_{C,i}$  than in  $\vec{S}_{C,i}^D$ . In other words, all objects will have higher membership degrees in  $\vec{S}_{C,i}$  of  $C$  than in  $\vec{S}_{C,i}^D$  of  $D$ . Thus, the concept defined by  $\vec{S}_{C,i}^D$  is a sub-concept of  $\vec{S}_{C,i}$  according to the definition 3 about concept subsumption. For all  $\vec{S}_{C,i} \in \widehat{C}$ , we know there is a sub-concept  $\vec{S}_{C,i}^D \in \widehat{D}$  for it, then  $D \subseteq C$ .

There are some implications in our  $f_{om}$ - $\mathcal{DL}$  as follows. Due to the lack of space, we omit the proofs of these implications.

- **Implication of Fuzzy Role:** If we have a fuzzy role instance  $\langle a, c \rangle^{v_1} \in R_1$ ,  $R_1 \subseteq R_2$ , then it implies that  $\langle a, c \rangle^{v_2} \in R_2$  and  $v_2 \geq v_1$ .
- **Implication of Fuzzy Concept:** For an object  $a$ , two fuzzy concepts  $C_1$  and  $C_2$ , if  $a^{v_1} \in C_1$ ,  $C_1 \subseteq C_2$ , then it implies that  $a^{v_2} \in C_2$  and  $v_2 \geq v_1$ .
- **Implication of Fuzzy Property** If object  $a$  possesses a fuzzy property  $P_1$  to a degree  $v_1$ , and  $P_1 \subseteq P_2$ , then it implies that  $a$  possesses  $P_2$  to a degree  $v_2$  where  $v_2$  is greater than or equal to  $v_1$ .

### 3.4 Satisfiability and Entailment

The notion of *satisfiability* of a fuzzy axiom  $E$  by a fuzzy interpretation  $\mathcal{I}$ , denoted  $\mathcal{I} \models E$ , is defined as follows:  $\mathcal{I} \models \langle \alpha \geq n \rangle$ , where  $\alpha$  is a concept or a role assertion axiom, if and only if  $\alpha^{\mathcal{I}} \geq n$ . Similarly, for the other relations such as  $\leq$ ,  $=$ ,  $<$ ,  $>$ . Moreover,  $\mathcal{I} \models a \approx b$  iff  $a^{\mathcal{I}} = b^{\mathcal{I}}$  and  $\mathcal{I} \models a \neq b$  iff  $a^{\mathcal{I}} \neq b^{\mathcal{I}}$ . For a set of fuzzy axioms  $\varepsilon$ , we say that  $\mathcal{I}$  *satisfies*  $\varepsilon$  iff  $\mathcal{I}$  satisfies each element in  $\varepsilon$ . If  $\mathcal{I} \models E$  (resp.  $\mathcal{I} \models \varepsilon$ ) we say that  $\mathcal{I}$  is a *model* of  $E$  (resp.  $\varepsilon$ ).  $\mathcal{I}$  *satisfies* (is a *model* of) a fuzzy knowledge base  $\mathcal{K} = \langle \mathcal{T}, \mathcal{R}, \mathcal{A} \rangle$ , denoted  $\mathcal{I} \models \mathcal{K}$ , iff  $\mathcal{I}$  is a model of each component  $\mathcal{T}$ ,  $\mathcal{R}$  and  $\mathcal{A}$ , respectively. A fuzzy axiom  $E$  is entailed by a knowledge base  $\mathcal{K}$ , denoted as  $\mathcal{K} \models E$  iff every model of  $\mathcal{K}$  satisfies  $E$ .

## 4 Object Membership Measurement

In this section, we introduce a mechanism to measure object membership. According to [10], membership degree of an object  $a$  in concept  $C$  depends on the comparison of the description of object  $a$  and characteristic vectors of the definition of  $C$ . If an object  $a$  possesses all the defining properties in one of characteristic vectors  $\vec{S}_i$  of  $C$  to a degree greater than zero, then  $a$  is a member of  $C$  to some degree.<sup>2</sup> Besides, while object  $a$  possesses all the defining properties of any  $\vec{S}_i$  of  $C$  to degrees which are greater than or equal to the minimum requirements of all defining properties of the specific  $\vec{S}_i$  in  $C$ , the membership of  $a$  in concept  $C$  is equal to one. For the reason that concepts are represented by N-properties and L-properties while objects are represented by fuzzy instance properties, and properties in our model may not be independent, we need to do property alignment (aligning fuzzy instance properties of objects to defining properties of concepts) before measuring the membership of objects in concepts based on properties comparison.

### 4.1 Measuring Degrees of Objects Possessing Defining Properties of Concepts

For the reason that a concept is defined by a set of disjoint characteristic vectors, we need to align the property vector of object  $a$  to each characteristic vectors. We define a function for the alignment between object description and characteristic vectors.

$$\text{alignO} : O_a \times S_x \rightarrow S_x^a$$

where  $P_a$  is the set of object descriptions,  $S_x$  is set of characteristic vectors and  $S_x^a$  is the set of aligned property vectors. The function  $\text{alignO}(\vec{O}_a, \vec{S}_x)$  is used to align object description  $\vec{O}_a$  to a characteristic vector  $\vec{S}_x$ , the result of  $\text{alignO}(\vec{O}_a, \vec{S}_x)$  is an aligned property vector  $\vec{S}_x^a$  as following:

$$\vec{S}_x^a = p_{x,1}(t_{x,1}^a) \cap p_{x,2}(t_{x,2}^a) \cap \dots \cap p_{x,n}(t_{x,n}^a), 1 \leq j \leq n$$

where  $n$  is the number of properties of  $\vec{S}_x$  and  $t_{x,j}^a$  is the degree of object  $a$  possessing property  $p_{x,j}$  in characteristic vector  $\vec{S}_x$ . In  $f_{om}\text{-}\mathcal{DL}$ , we can obtain the degree of object  $a$  possessing each defining property  $p_{x,j}$  ( $p_{x,j}$  can be an N-property or an L-property) based on the semantic axioms in table 1.

### 4.2 Calculation of Object Fuzzy Memberships in Concepts

For a concept  $C$  and object  $a$ , we can align  $\vec{p}_a$  to each characteristic vector  $\vec{S}_x$  of  $C$  and get its aligned property vector  $\vec{S}_x^a$ . The degree of a description of object  $a$  denoted by  $\vec{O}_a$  satisfying the minimum requirements of a characteristic vector  $\vec{S}_x$  is calculated by a comparison function of vectors.

$$\varphi : S_x^a \times S_x \rightarrow [0, 1]$$

<sup>2</sup> If object  $a$  possesses all the defining properties of  $\vec{S}_i$  of  $C$  to higher degrees, then its membership degree in  $C$  is higher.

where  $S_x^a$  is the set of aligned property vectors and  $S_x$  is the set of characteristic vectors. There are some axioms for  $\varphi(\vec{S}_x^a, \vec{S}_x)$  to observe.

**Axiom 1.** For a characteristic vector  $\vec{S}_x$  of a concept and its aligned property vector  $\vec{S}_x^a$ , if for some properties  $p_{x,i}$  in  $\vec{S}_x^a$ , we have  $t_{x,i}^a = 0$ , then  $\varphi(\vec{S}_x^a, \vec{S}_x) = 0$ .

**Axiom 2.** For a characteristic vector  $\vec{S}_x$  of a concept and its aligned property vector  $\vec{S}_x^a$ , if for each properties  $p_{x,i}$  in  $\vec{S}_x^a$ , we have  $t_{x,i}^a \geq w_{x,i}$ , then  $\varphi(\vec{S}_x^a, \vec{S}_x) = 1$ .

**Axiom 3.** For an object description  $\vec{O}_a$ , two characteristic vectors  $\vec{S}_{x1}$  and  $\vec{S}_{x2}$  of a concept,  $\vec{S}_{x1}^a$  is the aligned property vector of  $\vec{O}_a$  for  $\vec{S}_{x1}$  and  $\vec{S}_{x2}^a$  is the aligned property vector of  $\vec{O}_a$  for  $\vec{S}_{x2}$ , if  $w_{x1,i} \leq w_{x2,i}$  for some properties  $p_{x,i}$ , and  $w_{x1,j} = w_{x2,j}$  for others properties  $p_{x,j}$  where  $i \neq j$ , then  $\varphi(\vec{S}_{x1}^a, \vec{S}_{x1}) \geq \varphi(\vec{S}_{x2}^a, \vec{S}_{x2})$ .

**Axiom 4.** For a characteristic vector  $\vec{S}_x$  of a concept, two aligned property vectors  $\vec{S}_x^a$  and  $\vec{S}_x^b$  for object  $a$  and  $b$  respectively, if  $t_{x,i}^a \geq t_{x,i}^b$  for some properties  $p_{x,i}$  and  $t_{x,j}^a = t_{x,j}^b$  for others properties  $p_{x,j}$  where  $i \neq j$ , then  $\varphi(\vec{S}_x^a, \vec{S}_x) \geq \varphi(\vec{S}_x^b, \vec{S}_x)$ .

Axioms 1 and 2 specify the boundary cases of objects satisfying the minimum requirements of properties of concepts. Axioms 3 and 4 concern how the degree of an object property vector satisfying the minimum requirement of a characteristic vector is varied.

Here, we present a possible function which satisfies axioms 6, 7, 8 and 9.

$$\varphi(\vec{S}_x^a, \vec{S}_x) = \min(\tau_1, \tau_2, \dots, \tau_n) \tag{2}$$

where

$$\tau_i = \begin{cases} \frac{t_{x,i}^a}{w_{x,i}} & t_{x,i}^a < w_{x,i} \\ 1 & t_{x,i}^a \geq w_{x,i} \end{cases} \tag{3}$$

where  $t_{x,i}^a$  is the degree to which  $a$  possessing property  $p_{x,i}$  and  $w_{x,i}$  is the minimum requirement of property  $p_{x,i}$  in  $\vec{S}_x$ .

Besides, we consider the fuzzy membership of an object  $a$  in fuzzy concept  $C$  depends on the following equation:

$$\mu_C(a) = \max(\varphi(\vec{S}_1^a, \vec{S}_1), \varphi(\vec{S}_2^a, \vec{S}_2), \dots, \varphi(\vec{S}_n^a, \vec{S}_n)) \tag{4}$$

One object may satisfy all the property minimum requirements of more than one characteristic vectors. We choose the maximal value of  $\varphi(\vec{S}_i^a, \vec{S}_i)$  as the the membership of  $a$  in  $C$  because that the relation among  $\vec{S}_i$  is a disjunction. This is in line with fuzzy logic. Example 5 is used to illustrate the mechanism of measuring object membership based on object description and concept definition.

*Example 5.* For the concepts  $SC$  and a customer  $O_1$  in the examples 1 and 2. Firstly, we need to use *alignO* function to align the description of  $O_1$  for the concept definition  $SC$  as follow.

$$\vec{O}_1 = (\geq_5 \text{ buy.expensiveItem}(0.4) \sqcap \text{buy.expensiveItem}(0.8)) \sqcup (\geq_{100} \text{ buy.Item}(0.02) \sqcap \geq_1 \text{ buy.expensiveItem}(1))$$



The alignment is based on the interpreted functions of each N-property or L-property shown in table 1, i.e., the degrees of each object possessing defining properties (e.g., ‘ $\exists$ buy.expensiveItem’) is calculated based on all property members (e.g., ‘buy.Furniture00002’) possessed by the object for the corresponding property (e.g., ‘buy.expensiveItem’).<sup>3</sup> For example, customer  $O_1$  has property members for property ‘ $\exists$ buy.Item’ such as  $O_1$  possessing ‘buy.Furniture00002’ and ‘buy.Eproduct00307’, and these two property members belong to ‘ $\exists$ buy.Item’ to degree 1. Then the degree of object  $O_1$  possessing the property ‘ $\exists$ buy.Item’ is calculated using equation 1 as following:

$$(\exists\text{buy.Item})^{\mathcal{I}}(O_1) = \sup_{b \in \Delta^{\mathcal{I}}} \{ \min(R^{\mathcal{I}}(a, b), C^{\mathcal{I}}(b)) = \max(1, 1) = 1$$

We can then get the membership of  $O_1$  for  $SC$  by axioms 1-4 and equations 2, 3, 4 as following:

$$\begin{aligned} \varphi(\vec{S}_1^{O_1}, \vec{S}_1) &= 0.4, \varphi(\vec{S}_2^{O_1}, \vec{S}_2) = 0.02 \\ \mu_{SC}(O_1) &= \max(\varphi(\vec{S}_1^{O_1}, \vec{S}_1), \varphi(\vec{S}_2^{O_1}, \vec{S}_2)) = \max(0.4, 0.02) = 0.4 \end{aligned}$$

## 5 Discussion

There is no *property* element in previous fuzzy description logic. However, according to the study of cognitive psychology, properties are natural and widely used for representing concepts and objects by people in real life. In the proposed fuzzy description logic  $f_{om}\text{-}\mathcal{DL}$ , we introduce two kinds of properties which are L-Properties and N-properties. They are used to measure the degree an object possesses properties qualitatively and quantitatively, respectively. An L-property is a qualitative measurement of an object possessing a property based on aggregating all property members the object possessing for the property, while an N-property is a quantitative measurement of an object possessing a property based on a number restriction on all property members the object possessing for the property. To our best knowledge, there is no previous work which can express the qualitative measurement for the degree of an object possessing a property. These two measurements are frequently used measurements from two perspectives of people in reality.

In previous works, there lacks a formal mechanism to measure object membership in concepts based on object description and concept definition automatically, and object memberships in concepts are given manually or based on user defined fuzzy functions. Thus, the object membership determination are tedious, time-consuming and it depends on the user’s knowledge on the domain. However, psychologists find that object membership depends on how well an object satisfies the definition of a concept. That is a more reasonable and natural way to obtain object membership. In our work, according to the theories of cognitive psychology, we firstly define the syntax of concept definition and object description. Based on them, the mechanism we propose can measure object membership by matching properties in object description and defining properties in concept definition automatically. We believe it is a useful theoretical support to obtain the fuzzy membership of objects for concepts in fuzzy ontologies on Semantic Web. Especially, it could be a theoretical counterpart for the model presented by Cai and Leung in [10].

<sup>3</sup> For the interest of space, we omit the calculation details here.

## 6 Conclusion

In this paper, we propose a fuzzy description logic by combining the classical view in cognitive psychology and fuzzy set theory. In the proposed fuzzy description logic  $f_{om}\text{-}\mathcal{DL}$ , two kinds of properties named N-property and L-property are introduced. They are quantitative measurements and qualitative measurements of an object possessing a property, and previous works cannot express and represent the qualitative measurements of an object possessing a property. The subsumption of concepts, roles and properties are defined. We also present theories and implications about concept subsumption and property subsumption. A formal mechanism to measure object membership based on properties in object description and defining properties in concept definition automatically, while previous fuzzy description logic do not provide such a mechanism. For the reason that the proposed description logic can measure object membership automatically, we believe that it is useful to the Semantic Web community for reasoning the fuzzy membership of objects for concepts in fuzzy ontologies.

## References

1. Straccia, U.: Towards a fuzzy description logic for the semantic web. In: Gómez-Pérez, A., Euzenat, J. (eds.) *ESWC 2005*. LNCS, vol. 3532, pp. 167–181. Springer, Heidelberg (2005)
2. Straccia, U.: A fuzzy description logic. In: *AAAI 1998/IAAI 1998: Proceedings of the Fifteenth National/Tenth Conference on Artificial Intelligence/Innovative Applications of Artificial Intelligence*, pp. 594–599 (1998)
3. Stoilos, G., Stamou, G., Tzouvaras, V., Pan, J.Z., Horrocks, I.: The fuzzy description logic  $f\text{-}\mathcal{SH}\mathcal{LN}$ . In: *Proc. of the International Workshop on Uncertainty Reasoning for the Semantic Web (2005)*
4. Murphy, G.L.: *The big book of concepts*. MIT Press, Cambridge (2002)
5. Galotti, K.M.: *Cognitive Psychology In and Out of the Laboratory*, 3rd edn. Wadsworth, Belmont (2004)
6. Zadeh, L.A.: Fuzzy sets. *Information and Control* 8, 338–353 (1965)
7. Baader, F., Calvanese, D., McGuinness, D., Nardi, D., Schneider, P.P. (eds.): *The description logic handbook: theory, implementation, and applications*. Cambridge University Press, New York (2003)
8. Zadeh, L.A.: Fuzzy logic. *Computer* 21(4), 83–93 (1988)
9. Klir, G.J., Yuan, B.: *Fuzzy sets and fuzzy logic: theory and applications*. Prentice hall PTR, Englewood Cliffs (1995)
10. Cai, Y., Leung, H.F.: A formal model of fuzzy ontology with property hierarchy and object membership. In: Li, Q., Spaccapietra, S., Yu, E., Olivé, A. (eds.) *ER 2008*. LNCS, vol. 5231, pp. 69–82. Springer, Heidelberg (2008)
11. Smith, E.E., Medin, D.L.: *Categories and Concepts*. Harvard University Press, Cambridge (1981)
12. Yager, R.R.: On ordered weighted averaging aggregation operators in multicriteria decision-making. *IEEE Trans. Syst. Man Cybern.* 18(1), 183–190 (1988)
13. Yager, R.R.: On mean type aggregation. *IEEE Transactions on Systems, Man and Cybernetics* 26, 209–221 (1996)

# Decomposition-Based Optimization for Debugging of Inconsistent OWL DL Ontologies

Jianfeng Du<sup>1,2</sup> and Guilin Qi<sup>3</sup>

<sup>1</sup> Guangdong University of Foreign Studies, Guangzhou 510006, China  
jfd@mail.gdufs.edu.cn

<sup>2</sup> State Key Laboratory of Computer Science, Institute of Software,  
Chinese Academy of Sciences, Beijing 100190, China

<sup>3</sup> School of Computer Science and Engineering,  
Southeast University, Nanjing 211189, China

**Abstract.** Ontology debugging plays an important role in tackling inconsistency in OWL DL ontologies. It computes all minimal inconsistent sub-ontologies (MISs) of an inconsistent ontology. However, the computation of all MISs is costly. Existing module extraction methods that optimize the debugging of ontology entailments are not sufficient to optimize the computation of all MISs, because a module (i.e. a sub-ontology) that covers all MISs can be too large to be handled by existing OWL DL debugging facilities. In order to generate smaller sub-ontologies, we propose a novel method for computing a set of sub-ontologies from an inconsistent OWL DL ontology so that the computation of all MISs can be separately performed in each resulting sub-ontology and the union of computational results yields exactly the set of all MISs. Experimental results show that this method significantly improves the scalability for computing all MISs of an inconsistent ontology.

## 1 Introduction

OWL DL<sup>1</sup> is a standard language proposed by the W3C organization for modeling ontologies in the Semantic Web. It is a syntactic variant of Description Logic (DL) *SHOIN(D)* [5]. As *SHOIN(D)* is a fragment of First-Order Logic (FOL), *SHOIN(D)* (i.e. OWL DL) inherits from FOL the property *ex falso sequitur quodlibet* (from falsehood/contradiction follows anything), thus inconsistency in an OWL DL ontology is fatal. As inconsistency can be caused by unsatisfiable concepts, existing work pays much attention to debugging of unsatisfiable concepts in an ontology (e.g. [10,8]), which computes all minimal sub-ontologies in which a given concept is unsatisfiable. However debugging of unsatisfiable concepts does not directly deal with inconsistency. To handle inconsistency directly, Haase *et al.* [4] define the task of inconsistency debugging in an ontology as computing all minimal inconsistent sub-ontologies (MISs) of the ontology. MISs are very useful in tackling inconsistency because they contain erroneous axioms

---

<sup>1</sup> <http://www.w3.org/TR/owl-semantic/>

in an inconsistent ontology. Inconsistency can be tackled by fixing or removing axioms according to their appearance in MISs. Furthermore, MISs can be used to compute all maximal consistent sub-ontologies by using Reiter’s Hitting Set Tree (HST) algorithm [9], some of which can be used as repaired results for an inconsistent ontology.

Though MISs are useful in tackling inconsistency, they are computationally intensive due to the high complexity of consistency checking in OWL DL. To optimize the computation of all MISs, the current approach to optimizing ontology debugging, namely module extraction, can possibly be adapted. Existing module extraction methods [12,1] extract from a given ontology a sub-ontology that covers all minimal subsets of axioms responsible for a given ontology entailment. One may want to adapt existing module extraction approach to finding justifications to optimize the computation of all MISs. However, even if this approach is possible, it may not take effect when the union set of all MISs is very large. Take an ontology  $\mathcal{O}$  for example, which consists of the following axioms.

ChiefActress $\sqsubseteq$ Actress	Actress $\sqcap$ Actor $\sqsubseteq \perp$	ChiefActress( $a_1$ )	Actor( $a_1$ )
...	...	ChiefActress( $a_n$ )	Actor( $a_n$ )

It can be seen that the set of MISs of  $\mathcal{O}$  is  $\{\text{ChiefActress} \sqsubseteq \text{Actress}, \text{Actress} \sqcap \text{Actor} \sqsubseteq \perp, \text{ChiefActress}(a_i), \text{Actor}(a_i) \mid 1 \leq i \leq n\}$ . Hence a sub-ontology covering all these MISs is exactly  $\mathcal{O}$ . This example shows that module extraction is insufficient in optimizing the computation of all MISs. We therefore need a new method that can generate smaller sub-ontologies from which MISs are computed.

In order to make the computation of all MISs easier, we propose a new method for computing a set of sub-ontologies from which MISs can be separately computed (see Section 3). The method is based on a similar process as the one given in [1], i.e. compilation from OWL DL to Propositional Logic, but extends the compilation to yield a *labeled propositional program* in which every ground clause is associated with a label. The label stands for the axiom from which the corresponding ground clause is compiled. After compiling the labeled propositional program, our proposed method decomposes it into disjoint components and constructs sub-ontologies based on the labels in the resulting components. We prove that MISs can be separately computed from each resulting sub-ontology and the union of computational results is exactly the set of all MISs. For the aforementioned example, our proposed method computes from  $\mathcal{O}$  the set of sub-ontologies  $\{\text{ChiefActress} \sqsubseteq \text{Actress}, \text{Actress} \sqcap \text{Actor} \sqsubseteq \perp, \text{ChiefActress}(a_i), \text{Actor}(a_i) \mid 1 \leq i \leq n\}$ , each of which is exactly a MIS of  $\mathcal{O}$ .

We implement the proposed method as a prototype system and test on two groups of inconsistent ontologies (see Section 4). The ontologies in the first group are modified from existing incoherent ontologies. For two out of the three ontologies in this group, the computation of all MISs cannot be done in four hours by the well-known Pellet system [11] (simply Pellet), but can be computed in a few minutes by our system. The ontologies in the second group are modified from the University Benchmark (UOBM-Lite) [7] by inserting conflicts. Experimental results on this group show that our system significantly outperforms Pellet and scales well up to hundreds of conflicts when the number of conflicts increases.

## 2 Preliminaries

**OWL DL and Debugging.** OWL DL corresponds to DL *SHOIN* [5].<sup>2</sup> We assume that the reader is familiar with OWL DL and thus we do not describe it in detail, but recall that an OWL DL ontology  $\mathcal{O}$  consists of a set of axioms. These axioms include *terminological axioms* (i.e. *concept inclusion axioms*  $C \sqsubseteq D$ , *transitivity axioms*  $\text{Trans}(R)$  and *role inclusion axioms*  $R \sqsubseteq S$ ) and *assertional axioms* (i.e. *concept assertions*  $C(a)$ , *role assertions*  $R(a, b)$ , *equality assertions*  $a \approx b$  and *inequality assertions*  $a \not\approx b$ ), where  $C$  and  $D$  are OWL DL concepts,  $R$  and  $S$  roles, and  $a$  and  $b$  individuals.

We briefly introduce the direct model-theoretic semantics for an OWL DL ontology  $\mathcal{O}$ . An *interpretation*  $\mathcal{I} = (\Delta^{\mathcal{I}}, \cdot^{\mathcal{I}})$  consists of a *domain*  $\Delta^{\mathcal{I}}$  and a function  $\cdot^{\mathcal{I}}$  that maps every atomic concept  $A$  to a set  $A^{\mathcal{I}} \subseteq \Delta^{\mathcal{I}}$ , every atomic role  $R$  to a binary relation  $R^{\mathcal{I}} \subseteq \Delta^{\mathcal{I}} \times \Delta^{\mathcal{I}}$ , and every individual  $a$  to  $a^{\mathcal{I}} \in \Delta^{\mathcal{I}}$ .  $\mathcal{I}$  is called a *model* of  $\mathcal{O}$  if every axiom in  $\mathcal{O}$  is satisfied by  $\mathcal{I}$ .  $\mathcal{O}$  is said to be *consistent* if it has a model. A concept  $C$  is said to be *satisfiable* in  $\mathcal{O}$  if there exists a model  $\mathcal{I}$  of  $\mathcal{O}$  such that  $C^{\mathcal{I}} \neq \emptyset$ .  $\mathcal{O}$  is said to be *coherent* if every atomic concept in it is satisfiable.

A *sub-ontology*  $\mathcal{O}'$  of  $\mathcal{O}$  is a subset of axioms in  $\mathcal{O}$ . A *minimal inconsistent sub-ontology* (MIS)  $\mathcal{O}'$  of  $\mathcal{O}$  is an inconsistent sub-ontology of  $\mathcal{O}$  that has not any proper subset  $\mathcal{O}''$  such that  $\mathcal{O}''$  is also inconsistent. The task of debugging of inconsistency in an inconsistent ontology  $\mathcal{O}$  is to compute all MISs of  $\mathcal{O}$  [4].

**First-order Logic and Axiomatizing Equality.** *Terms* are variables, constants or functional terms of the form  $f(t_1, \dots, t_n)$ , where  $f$  is a function symbol of arity  $n$  and  $t_1, \dots, t_n$  are terms.<sup>3</sup> We only consider unary function symbols because only unary function symbols occur in first-order logic programs that are translated from OWL DL ontologies. *Atoms* are of the form  $T(t_1, \dots, t_n)$  where  $T$  is a predicate symbol of arity  $n$  and  $t_1, \dots, t_n$  are terms. A *literal* is a positive or negative atom and a *clause* is a disjunction of literals. Terms, atoms and clauses that do not contain variables are called *ground*.

A *first-order logic program* (FOL program) is a set of clauses in which all variables are universally quantified. For a clause  $cl = \neg A_1 \vee \dots \vee \neg A_n \vee B_1 \vee \dots \vee B_m$ , the set of atoms  $\{A_1, \dots, A_n\}$  is denoted by  $cl^-$ , whereas the set of atoms  $\{B_1, \dots, B_m\}$  is denoted by  $cl^+$ . By  $|S|$  we denote the cardinality of a set  $S$ . A clause  $cl$  is called a *fact* if  $|cl^-| = 0$ .

A *propositional program*  $\Pi$  is a FOL program consisting of only ground clauses. The set of ground atoms occurring in  $\Pi$  is denoted by  $\text{atoms}(\Pi)$ .

For a FOL program  $P$ , the set of ground terms (resp. ground atoms) defined from the first-order signature of  $P$  is called the *Herbrand universe* (resp. *Herbrand base*) of  $P$ , denoted by  $\text{HU}(P)$  (resp.  $\text{HB}(P)$ ). The set of ground clauses

<sup>2</sup> OWL DL also involves datatypes which can be handled by our method with some extensions. But we do not complicate our presentation by considering them here.

<sup>3</sup> Throughout this paper, we use (possibly with subscripts)  $x, y, z$  for variables,  $a, b, c$  for constants, and  $s, t$  for terms.

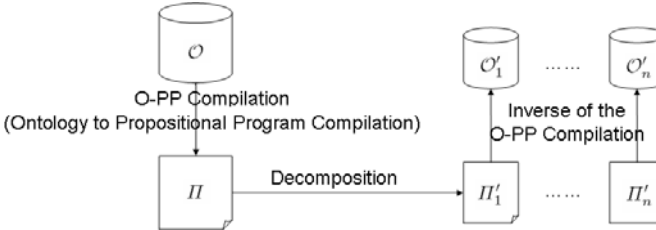
obtained by replacing all variables occurring in each clause in  $P$  with ground terms from  $\text{HU}(P)$  is called the *primary grounding* of  $P$ , denoted by  $\mathcal{G}(P)$ . An *interpretation*  $M$  of  $P$  is a set of ground atoms in  $\text{HB}(P)$ ; it is a *model* of  $P$  if for any ground clause  $cl \in \mathcal{G}(P)$  such that  $cl^- \subseteq M$ ,  $cl^+ \cap M \neq \emptyset$ ; it is a *minimal model* of  $P$  if there is no model  $M'$  of  $P$  such that  $M' \subset M$ .  $P$  is said to be *satisfiable* if  $P$  has a model.

The FOL program  $P$  translated from an OWL DL ontology may contain the *equality predicate*  $\approx$ , which is interpreted as a *congruence relation* and different from ordinary predicates. This difference is not captured by the above first-order semantics. However, the equality predicate  $\approx$  can be explicitly axiomatized via a well-known transformation from [3]. Let  $\mathcal{E}(P)$  denote the FOL program consisting of the following clauses: (1)  $t \approx t$ , for each ground term  $t \in \text{HU}(P)$ ; (2)  $\neg(x \approx y) \vee y \approx x$ ; (3)  $\neg(x \approx y) \vee \neg(y \approx z) \vee x \approx z$ ; (4)  $\neg(x \approx y) \vee f(x) \approx f(y)$ , for each function symbol  $f$  occurring in  $P$ ; (5)  $\neg T(x_1, \dots, x_i, \dots, x_n) \vee \neg(x_i \approx y_i) \vee T(x_1, \dots, y_i, \dots, x_n)$ , for each predicate symbol  $T$  other than  $\approx$  occurring in  $P$  and each position  $i$ . Appending  $\mathcal{E}(P)$  to  $P$  allows to treat  $\approx$  as an ordinary predicate, i.e.,  $M$  is a model of  $P$  that interprets  $\approx$  as a congruence relation, iff for any ground clause  $cl \in \mathcal{G}(P \cup \mathcal{E}(P))$  such that  $cl^- \subseteq M$ ,  $cl^+ \cap M \neq \emptyset$ .

### 3 Finding Sub-ontologies to Optimize MIS Computation

Given an OWL DL ontology  $\mathcal{O}$ , we intend to compute a set of sub-ontologies  $\{\mathcal{O}'_i\}_{1 \leq i \leq n}$  of  $\mathcal{O}$  such that the union of sets of MISs of these sub-ontologies yields the set of all MISs of  $\mathcal{O}$ , i.e.,  $\bigcup_{i=1}^n \mathcal{M}_i$  is the set of MISs of  $\mathcal{O}$ , where  $\mathcal{M}_i$  is the set of MISs of  $\mathcal{O}'_i$  for  $1 \leq i \leq n$ . The direct computation of these sub-ontologies from  $\mathcal{O}$  is hard as we need to consider the internal of ontology reasoning algorithms. Hence we follow the idea given in our previous work [1] and consider compiling  $\mathcal{O}$  into a propositional program. It is easier to analyze propositional reasoning algorithms because a model of a propositional program is a subset of ground atoms occurring in it. That is, it is easier to analyze the compiled propositional program and use it to compute sub-ontologies of  $\mathcal{O}$  that preserve MISs.

The idea exploited in the current work is depicted in the following figure.



First, the given ontology  $\mathcal{O}$  is compiled to a propositional program  $\Pi$  using a so-called O-PP (Ontology to Propositional Program) compilation process. Then,  $\Pi$  is decomposed into a set of disjoint components  $\Pi'_1, \dots, \Pi'_n$ . Finally, a set of sub-ontologies  $\{\mathcal{O}'_i\}_{1 \leq i \leq n}$  of  $\mathcal{O}$  is extracted from  $\{\Pi'_i\}_{1 \leq i \leq n}$  respectively using the inverse process of the O-PP compilation. Suppose  $\mathcal{M}_i$  is the set of MISs

computed from  $\mathcal{O}'_i$  by applying existing OWL DL debugging facilities ( $1 \leq i \leq n$ ). Then  $\bigcup_{i=1}^n \mathcal{M}_i$  is the set of MISs of  $\mathcal{O}$ .

There are two key points in the proposed method. The first one is how to define the O-PP compilation process and its inverse process. The second one is how to define the decomposition process. We will describe our solutions to these key points in the following subsections.

### 3.1 The O-PP Compilation Process and Its Inverse Process

As *SHOIN* is a fragment of FOL, an OWL DL (i.e. *SHOIN*) ontology can be translated to a FOL program using standard translation methods. Then the FOL program can be further transformed to a propositional program using standard grounding methods. Thus an OWL DL ontology can be compiled to a propositional program using standard methods. However, the standard methods are insufficient to facilitate the inverse process because any manipulation of the final propositional program is hardly mapped to the original ontology.

To make the inverse process easy, we develop a method for translating an OWL DL ontology  $\mathcal{O}$  to a *labeled FOL program*  $\mathcal{L}(\mathcal{O})$ , which consists of *labeled clauses* of the form  $(cl, ax)$ , where  $ax$  is the axiom from which the clause  $cl$  is translated;  $ax$  is the empty label  $\diamond$  if  $cl$  is a clause used to axiomatize equality. The labeled FOL program  $\mathcal{L}(\mathcal{O})$  is defined through a translation from an axiom  $ax$  to a set  $\mathcal{L}(ax)$  of labeled clauses.  $\mathcal{L}(ax)$  is defined recursively in the following frame, where  $A$  and  $B$  are atomic concepts or  $\top$ ,  $a$  and  $b$  are individuals,  $C, C_1, C_2$  and  $D$  are *SHOIN* concepts,  $R$  and  $S$  are roles, and  $Q_X$  is  $X$  if  $X$  is an atomic concept, or a new globally unique concept name for  $X$  otherwise;  $\text{NNF}(C)$  denotes the result of translating  $C$  to the negation normal form;  $\text{ar}(R, s, t)$  denotes  $S(t, s)$  if  $R = S^-$  for some role name  $S$ , or  $R(s, t)$  otherwise.

$\begin{aligned} &\text{if } ax \equiv R \sqsubseteq S, \mathcal{L}(ax) = \{(\neg \text{ar}(R, x, y) \vee \text{ar}(S, x, y), ax)\}; \\ &\text{if } ax \equiv \text{Trans}(R), \mathcal{L}(ax) = \{(\neg \text{ar}(R, x, y) \vee \neg \text{ar}(R, y, z) \vee \text{ar}(R, x, z), ax)\}; \\ &\text{if } ax \equiv C \sqsubseteq D, \mathcal{L}(ax) = \{(cl, ax) \mid cl \in \mathcal{F}(\top \sqsubseteq \text{NNF}(\neg C \sqcup D))\}; \\ &\text{if } ax \equiv C(a), \mathcal{L}(ax) = \{(Q_D(a), ax)\} \cup \{(cl, ax) \mid cl \in \mathcal{F}(Q_D \sqsubseteq D)\} \text{ for } D = \text{NNF}(C); \\ &\text{if } ax \equiv R(a, b), \mathcal{L}(ax) = \{(\text{ar}(R, a, b), ax)\}; \\ &\text{if } ax \equiv a \approx b, \mathcal{L}(ax) = \{(a \approx b, ax)\}; \quad \text{if } ax \equiv a \not\approx b, \mathcal{L}(ax) = \{(\neg(a \approx b), ax)\}; \\ &\mathcal{F}(A \sqsubseteq B) = \{\neg A(x) \vee B(x)\}; \quad \mathcal{F}(A \sqsubseteq \neg B) = \{\neg A(x) \vee \neg B(x)\}; \\ &\mathcal{F}(A \sqsubseteq \{a\}) = \{\neg A(x) \vee x \approx a\}; \quad \mathcal{F}(A \sqsubseteq \neg\{a\}) = \{\neg A(a)\}; \\ &\mathcal{F}(A \sqsubseteq C_1 \sqcap C_2) = \mathcal{F}(A \sqsubseteq C_1) \cup \mathcal{F}(A \sqsubseteq C_2); \\ &\mathcal{F}(A \sqsubseteq C_1 \sqcup C_2) = \{\neg A(x) \vee Q_{C_1}(x) \vee Q_{C_2}(x)\} \cup \mathcal{F}(Q_{C_1} \sqsubseteq C_1) \cup \mathcal{F}(Q_{C_2} \sqsubseteq C_2); \\ &\mathcal{F}(A \sqsubseteq \exists R.C) = \{\neg A(x) \vee \text{ar}(R, x, f(x)), \neg A(x) \vee Q_C(f(x))\} \cup \mathcal{F}(Q_C \sqsubseteq C); \\ &\mathcal{F}(A \sqsubseteq \forall R.C) = \{\neg A(x) \vee \neg \text{ar}(R, x, y) \vee Q_C(y)\} \cup \mathcal{F}(Q_C \sqsubseteq C); \\ &\mathcal{F}(A \sqsubseteq_{\geq n} R) = \{\neg A(x) \vee \text{ar}(R, x, f_i(x)) \mid 1 \leq i \leq n\} \cup \\ &\quad \{\neg(f_j(x) \approx f_k(x)) \mid 1 \leq j < k \leq n\}; \\ &\mathcal{F}(A \sqsubseteq_{\leq n} R) = \{\neg A(x) \vee \bigvee_{i=1}^{n+1} \neg \text{ar}(R, x, y_i) \vee \bigvee_{j=1}^n \bigvee_{k=j+1}^{n+1} y_j \approx y_k\}. \end{aligned}$
--

Let  $\mathcal{L}^*(\mathcal{O})$  denote  $\bigcup_{ax \in \mathcal{O}} \mathcal{L}(ax)$ . Then  $\mathcal{L}(\mathcal{O})$  is defined as  $\mathcal{L}^*(\mathcal{O}) \cup \{(cl, \diamond) \mid cl \in \mathcal{E}(\{cl \mid (cl, ax) \in \mathcal{L}^*(\mathcal{O})\})\}$  if some equational atom  $s \approx t$  occurs positively in  $\mathcal{L}^*(\mathcal{O})$ , or defined as  $\mathcal{L}^*(\mathcal{O})$  otherwise. Recall that  $\mathcal{E}(P)$  for a FOL program  $P$  is used to axiomatize equality in  $P$ , as described in Section 2. Such translation from

$\mathcal{O}$  to  $\mathcal{L}(\mathcal{O})$  is based on the well-known processes of *structural transformation* [6], clausification and equality axiomatization.

*Example 1.* Our running example is adapted from the example given in Section 1 by adding  $ax_2, ax_3$  and  $ax_5$  and fixing  $n = 2$ . Let  $\mathcal{O} = \{ax_1, \dots, ax_9\}$ .

$$\begin{aligned} ax_1 &\equiv \text{ChiefActress} \sqsubseteq \text{Actress} & ax_2 &\equiv \text{ChiefActress} \sqcap \text{Man} \sqsubseteq \perp \\ ax_3 &\equiv \text{Actress} \sqsubseteq \text{Person} & ax_4 &\equiv \text{Actress} \sqcap \text{Actor} \sqsubseteq \perp \\ ax_5 &\equiv \text{Person} \sqsubseteq \text{Man} \sqcup \text{Woman} & ax_6 &\equiv \text{Actor}(a_1) & ax_7 &\equiv \text{Actor}(a_2) \\ ax_8 &\equiv \text{ChiefActress}(a_1) & ax_9 &\equiv \text{ChiefActress}(a_2) \end{aligned}$$

Then  $\mathcal{L}(\mathcal{O})$  consists of the following labeled clauses.

$$\begin{aligned} lcl_1 &\equiv (\neg \text{ChiefActress}(x) \vee \text{Actress}(x), ax_1) & lcl_2 &\equiv (\neg \text{ChiefActress}(x) \vee \neg \text{Man}(x), ax_2) \\ lcl_3 &\equiv (\neg \text{Actress}(x) \vee \text{Person}(x), ax_3) & lcl_4 &\equiv (\neg \text{Actress}(x) \vee \neg \text{Actor}(x), ax_4) \\ lcl_5 &\equiv (\neg \text{Person}(x) \vee \text{Man}(x) \vee \text{Woman}(x), ax_5) & lcl_6 &\equiv (\text{Actor}(a_1), ax_6) \\ lcl_7 &\equiv (\text{Actor}(a_2), ax_7) & lcl_8 &\equiv (\text{ChiefActress}(a_1), ax_8) & lcl_9 &\equiv (\text{ChiefActress}(a_2), ax_9) \end{aligned}$$

For a labeled FOL program  $P$ , by  $\text{Cl}(P)$  and  $\text{Ax}(P)$  we respectively denote the clause part and axiom part of  $P$ , i.e.,  $\text{Cl}(P) = \{cl \mid (cl, ax) \in P\}$  and  $\text{Ax}(P) = \{ax \in \mathcal{O} \mid (cl, ax) \in P\}$ . For a sub-ontology  $\mathcal{O}'$  of  $\mathcal{O}$ , by  $\rho(\mathcal{L}(\mathcal{O}), \mathcal{O}')$  we denote  $\{(cl, ax) \in \mathcal{L}(\mathcal{O}) \mid ax \in \{\diamond\} \cup \mathcal{O}'\}$ , which is the subset of  $\mathcal{L}(\mathcal{O})$  consisting of clauses translated from  $\mathcal{O}'$  and clauses used to axiomatize equality. The following lemma shows a relationship between  $\mathcal{O}'$  and  $\rho(\mathcal{L}(\mathcal{O}), \mathcal{O}')$ , which follows from the first-order semantics of OWL DL [6] and the fact that structural transformation and clausification preserve satisfiability.

**Lemma 1.** *For any subset  $\mathcal{O}'$  of  $\mathcal{O}$ ,  $\mathcal{O}'$  is consistent iff  $\text{Cl}(\rho(\mathcal{L}(\mathcal{O}), \mathcal{O}'))$  is satisfiable.  $\square$*

To complete the O-PP compilation process, we adapt the notion of bottom-up grounding given in [1] to a labeled FOL program  $P$  and define the *bottom-up grounding* of  $P$ , denoted by  $\mathcal{G}_{bu}(P)$ , as the least fixpoint of  $\Pi^{(n)}$  such that  $\Pi^{(0)} = \emptyset$  and for  $n \geq 1$ ,  $\Pi^{(n)} = \{(cl\sigma, ax) \mid (cl, ax) \in P, \sigma \text{ is a ground substitution such that } cl^{-\sigma} \subseteq \text{atoms}(\text{Cl}(\Pi^{(n-1)})) \text{ and } cl^{+\sigma} \subseteq \text{HB}(\text{Cl}(P))\}$ . Note that a propositional program is a special FOL program, so we also use  $\text{Cl}(\Pi)$  and  $\text{Ax}(\Pi)$  to respectively denote the clause part and axiom part of a labeled propositional program  $\Pi$ . This notion of bottom-up grounding only differs from the original one given in [1] on considering labels. The following lemma shows a relation between  $P$  and  $\mathcal{G}_{bu}(P)$ , proved analogously as in Lemma 3 of [1].

**Lemma 2.** *For a labeled FOL program  $P$  where the equality predicate has been axiomatized,  $\text{Cl}(\mathcal{G}_{bu}(P))$  has the same set of minimal models as  $\text{Cl}(P)$  has.  $\square$*

Note that  $\mathcal{G}_{bu}(\mathcal{L}(\mathcal{O}))$  can be infinite due to function symbols in  $\mathcal{L}(\mathcal{O})$  introduced by *skolemization* in the course of clausification. We will present an approximate grounding method to tackle this problem in Subsection 3.3. But here we first consider an O-PP compilation process which compiles  $\mathcal{O}$  to  $\mathcal{G}_{bu}(\mathcal{L}(\mathcal{O}))$ .

For a subset  $\Pi$  of  $\mathcal{G}_{bu}(\mathcal{L}(\mathcal{O}))$ , we can naturally define the inverse process of the O-PP compilation (on  $\Pi$ ) as computing  $\text{Ax}(\Pi)$ . Clearly this inverse process is easy; this explains why we introduce labels to the O-PP compilation process. The following lemma shows a relation between  $\text{Cl}(\Pi)$  and  $\text{Ax}(\Pi)$ .



**Lemma 3.** *For any subset  $\Pi$  of  $\mathcal{G}_{bu}(\mathcal{L}(\mathcal{O}))$ ,  $\text{Ax}(\Pi)$  is inconsistent if  $\text{Cl}(\Pi)$  is unsatisfiable.*

*Proof.* Since each clause in  $\text{Cl}(\Pi)$  is instantiated from  $\text{Cl}(\rho(\mathcal{L}(\mathcal{O}), \text{Ax}(\Pi)))$ ,  $\text{Cl}(\Pi)$  is a subset of  $\mathcal{G}(\text{Cl}(\rho(\mathcal{L}(\mathcal{O}), \text{Ax}(\Pi))))$ , i.e. the primary grounding of  $\text{Cl}(\rho(\mathcal{L}(\mathcal{O}), \text{Ax}(\Pi)))$ . When  $\text{Cl}(\Pi)$  is unsatisfiable,  $\mathcal{G}(\text{Cl}(\rho(\mathcal{L}(\mathcal{O}), \text{Ax}(\Pi))))$  is also unsatisfiable and so is  $\text{Cl}(\rho(\mathcal{L}(\mathcal{O}), \text{Ax}(\Pi)))$ . By Lemma 1,  $\text{Ax}(\Pi)$  is inconsistent.  $\square$

Lemma 3 implies that some MISs of  $\mathcal{O}$  can be computed from the axiom part of a subset  $\Pi$  of  $\mathcal{G}_{bu}(\mathcal{L}(\mathcal{O}))$  such that  $\text{Cl}(\Pi)$  is unsatisfiable. Hence we can first compute all small subsets of  $\mathcal{G}_{bu}(\mathcal{L}(\mathcal{O}))$  whose clause parts are unsatisfiable, then compute MISs of  $\mathcal{O}$  from them. The computation of such subsets of  $\mathcal{G}_{bu}(\mathcal{L}(\mathcal{O}))$  can be realized by decomposing  $\mathcal{G}_{bu}(\mathcal{L}(\mathcal{O}))$ , as shown in the next subsection.

### 3.2 The Decomposition Process

The basic idea of decomposing  $\mathcal{G}_{bu}(\mathcal{L}(\mathcal{O}))$  is to first remove a *satisfiable core* from  $\mathcal{G}_{bu}(\mathcal{L}(\mathcal{O}))$ , then compute *maximal connected components* of the remaining subset, where a subset  $\Pi_{sc}$  of  $\mathcal{G}_{bu}(\mathcal{L}(\mathcal{O}))$  is called a *satisfiable core* of  $\mathcal{G}_{bu}(\mathcal{L}(\mathcal{O}))$  if for any subset  $\mathcal{O}'$  of  $\mathcal{O}$ ,  $\text{Cl}(\mathcal{G}_{bu}(\rho(\mathcal{L}(\mathcal{O}), \mathcal{O}')) \cap \Pi_{sc})$  has a model  $M_0$  such that for all models  $M$  of  $\text{Cl}(\mathcal{G}_{bu}(\rho(\mathcal{L}(\mathcal{O}), \mathcal{O}')) \setminus \Pi_{sc})$ ,  $M_0 \cup M$  is a model of  $\text{Cl}(\mathcal{G}_{bu}(\rho(\mathcal{L}(\mathcal{O}), \mathcal{O}')))$ . This idea is similar with the one given in [1], which is used to compute a sub-ontology that preserves all minimal subsets of axioms responsible for a given ontology entailment. Here the notion of maximal connected component is adapted from the homonymous notion given in [1]. That is, a *connected component* of a labeled propositional program  $\Pi$  is a subset of  $\Pi$  such that any two clauses in its clause part are connected through common ground atoms, while a *maximal connected component* of  $\Pi$  is a connected component of  $\Pi$  which cannot be a proper subset of any connected component of  $\Pi$ . The following theorem shows the correctness of a method exploiting the above idea.

**Theorem 1.** *Let  $\Pi_{sc1}$  be a subset of  $\mathcal{G}_{bu}(\mathcal{L}(\mathcal{O}))$  such that for all  $(cl, ax) \in \Pi_{sc1}$ ,  $cl^+ \not\subseteq \text{atoms}(\text{Cl}(\mathcal{G}_{bu}(\mathcal{L}(\mathcal{O})) \setminus \Pi_{sc1}))$ . Let  $\Pi_{sc2}$  be a subset of  $\mathcal{G}_{bu}(\mathcal{L}(\mathcal{O})) \setminus \Pi_{sc1}$  such that for all  $(cl, ax) \in \Pi_{sc2}$ ,  $cl^- \not\subseteq \bigcup_{(cl, ax) \in \mathcal{G}_{bu}(\mathcal{L}(\mathcal{O})) \setminus (\Pi_{sc1} \cup \Pi_{sc2})} cl^+$ . Let  $\{\Pi_i\}_{1 \leq i \leq m}$  be the set of maximal connected components of  $\mathcal{G}_{bu}(\mathcal{L}(\mathcal{O})) \setminus (\Pi_{sc1} \cup \Pi_{sc2})$ . Let  $\mathcal{M}_i$  be the set of MISs of  $\text{Ax}(\Pi_i)$  for  $1 \leq i \leq m$ . Then  $\bigcup_{i=1}^m \mathcal{M}_i$  is the set of all MISs of  $\mathcal{O}$ .*

*Proof.* (i) It is clear that a MIS of  $\text{Ax}(\Pi_i)$  for any  $1 \leq i \leq m$  is a MIS of  $\mathcal{O}$ . (ii) Consider a MIS  $\mathcal{O}'$  of  $\mathcal{O}$ . By Lemma 1,  $\text{Cl}(\rho(\mathcal{L}(\mathcal{O}), \mathcal{O}'))$  is unsatisfiable. Let  $\Pi' = \mathcal{G}_{bu}(\rho(\mathcal{L}(\mathcal{O}), \mathcal{O}'))$ . By Lemma 2,  $\text{Cl}(\Pi')$  is also unsatisfiable. Let  $\Pi'_i = \Pi' \cap \Pi_i$  for all  $1 \leq i \leq m$ . Suppose  $\text{Cl}(\Pi'_i)$  is satisfiable for all  $1 \leq i \leq m$ . Let  $M_i$  be a model of  $\text{Cl}(\Pi'_i)$  for  $1 \leq i \leq m$  and  $M_0 = \text{atoms}(\text{Cl}(\Pi')) \cap \bigcup_{(cl, ax) \in \Pi_{sc1}} cl^+ \setminus \text{atoms}(\text{Cl}(\mathcal{G}_{bu}(\mathcal{L}(\mathcal{O})) \setminus \Pi_{sc1}))$ . Let  $\Pi'_0 = \Pi' \setminus \bigcup_{i=1}^m \Pi'_i$ . Then  $\Pi'_0 \subseteq \Pi_{sc1} \cup \Pi_{sc2}$ . For all  $cl \in \text{Cl}(\Pi_{sc1} \cap \Pi'_0)$ , since  $cl^+ \not\subseteq \text{atoms}(\text{Cl}(\mathcal{G}_{bu}(\mathcal{L}(\mathcal{O})) \setminus \Pi_{sc1}))$ , we have  $\emptyset \subset cl^+ \setminus \text{atoms}(\text{Cl}(\mathcal{G}_{bu}(\mathcal{L}(\mathcal{O})) \setminus \Pi_{sc1})) \subseteq M_0$ , thus  $M_0 \cap cl^+ \neq \emptyset$ . For all  $cl \in \text{Cl}(\Pi_{sc2} \cap \Pi'_0)$ , since  $M_0 \subseteq \bigcup_{(cl, ax) \in \mathcal{G}_{bu}(\mathcal{L}(\mathcal{O})) \setminus (\Pi_{sc1} \cup \Pi_{sc2})} cl^+$ , we have  $cl^- \not\subseteq M_0$ . It follows that

**Algorithm 1.** Decompose( $\Pi$ )

1.  $\Pi_{sc1} := \Pi$ ;  $\Pi'_{sc1} := \emptyset$ ;
2. **while**  $\Pi'_{sc1} \neq \Pi_{sc1}$  **do**
3.    $\Pi'_{sc1} := \Pi_{sc1}$ ;
4.   **for each**  $(cl, ax) \in \Pi_{sc1}$  s.t.  $cl^+ \subseteq \text{atoms}(\text{Cl}(\Pi \setminus \Pi_{sc1}))$  **do**
5.      $\Pi_{sc1} := \Pi_{sc1} \setminus \{(cl, ax)\}$ ;
6.  $\Pi_{sc2} := \emptyset$ ;  $\Pi'_{sc2} := \Pi$ ;
7. **while**  $\Pi'_{sc2} \neq \Pi_{sc2}$  **do**
8.    $\Pi'_{sc2} := \Pi_{sc2}$ ;
9.   **for each**  $(cl, ax) \in \Pi \setminus (\Pi_{sc1} \cup \Pi_{sc2})$  s.t.  $cl^- \not\subseteq \bigcup_{(cl, ax) \in \Pi \setminus (\Pi_{sc1} \cup \Pi_{sc2})} cl^+$  **do**
10.     $\Pi_{sc2} := \Pi_{sc2} \cup \{(cl, ax)\}$ ;
11. **return** the set of maximal connected components of  $\Pi \setminus (\Pi_{sc1} \cup \Pi_{sc2})$ ;

$M_0$  is a model of  $\text{Cl}(\Pi'_0)$ . Since  $\bigcup_{i=1}^m M_i \subseteq \bigcup_{(cl, ax) \in \mathcal{G}_{bu}(\mathcal{L}(\mathcal{O})) \setminus (\Pi_{sc1} \cup \Pi_{sc2})} cl^+$  and for all  $0 \leq i \leq m$  and  $1 \leq j \leq m$ ,  $M_i \cap \text{atoms}(\Pi'_j) = \emptyset$  if  $i \neq j$ ,  $\bigcup_{i=0}^m M_i$  is a model of  $\bigcup_{i=0}^m \text{Cl}(\Pi'_i) = \text{Cl}(\Pi')$ , contradicting that  $\text{Cl}(\Pi')$  is unsatisfiable. Hence  $\text{Cl}(\Pi'_k)$  is unsatisfiable for some  $1 \leq k \leq m$ . Since  $\Pi'_k$  is a subset of  $\mathcal{G}_{bu}(\mathcal{L}(\mathcal{O}))$ , by Lemma 3,  $\text{Ax}(\Pi'_k)$  is inconsistent. Since  $\text{Ax}(\Pi'_k) \subseteq \text{Ax}(\Pi') = \mathcal{O}'$  and  $\mathcal{O}'$  is a MIS of  $\mathcal{O}$ , we have  $\mathcal{O}' = \text{Ax}(\Pi'_k) \subseteq \text{Ax}(\Pi_k)$ , thus  $\mathcal{O}' \in \mathcal{M}_k$ .  $\square$

Algorithm 1 shows how to decompose  $\mathcal{G}_{bu}(\mathcal{L}(\mathcal{O}))$  (where the input  $\Pi$  is set as  $\mathcal{G}_{bu}(\mathcal{L}(\mathcal{O}))$ ) using the method given in Theorem 1. The set of maximal connected components of  $\Pi \setminus (\Pi_{sc1} \cup \Pi_{sc2})$  returned by the algorithm can be efficiently computed by the well-known union-find algorithm (cf. [http://en.wikipedia.org/wiki/Union-find\\_algorithm](http://en.wikipedia.org/wiki/Union-find_algorithm)). It can be seen that  $\Pi_{sc1}$  and  $\Pi_{sc2}$  computed in  $\text{Decompose}(\mathcal{G}_{bu}(\mathcal{L}(\mathcal{O})))$  satisfy the conditions of the homonymous subsets of  $\mathcal{G}_{bu}(\mathcal{L}(\mathcal{O}))$  given in Theorem 1, thus we have the following theorem.

**Theorem 2.** Let  $\{\Pi_i\}_{1 \leq i \leq m}$  be returned by  $\text{Decompose}(\mathcal{G}_{bu}(\mathcal{L}(\mathcal{O})))$ . Let  $\mathcal{M}_i$  be the set of MISs of  $\text{Ax}(\Pi_i)$ . Then  $\bigcup_{i=1}^m \mathcal{M}_i$  is the set of all MISs of  $\mathcal{O}$ .  $\square$

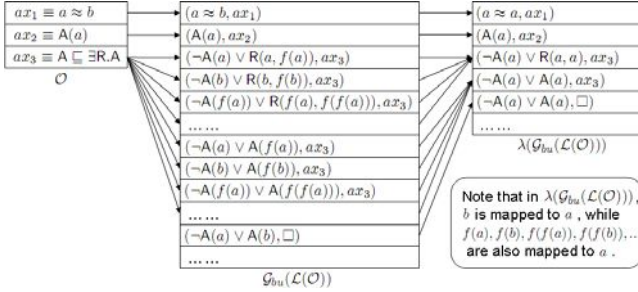
*Example 2.* Continue Example 1. By applying a fixpoint-evaluation manner, we can compute that  $\mathcal{G}_{bu}(\mathcal{L}(\mathcal{O})) = \{lcl'_1, \dots, lcl'_{14}\}$ . (Note that the labeled clauses are ordered to save space.)

$$\begin{aligned}
lcl'_1 &\equiv (\neg \text{ChiefActress}(a_1) \vee \text{Actress}(a_1), ax_1) & lcl'_2 &\equiv (\neg \text{ChiefActress}(a_1) \vee \neg \text{Man}(a_1), ax_2) \\
lcl'_3 &\equiv (\neg \text{Actress}(a_1) \vee \text{Person}(a_1), ax_3) & lcl'_4 &\equiv (\neg \text{Actress}(a_1) \vee \neg \text{Actor}(a_1), ax_4) \\
lcl'_5 &\equiv (\neg \text{Person}(a_1) \vee \text{Man}(a_1) \vee \text{Woman}(a_1), ax_5) & lcl'_6 &\equiv (\text{Actor}(a_1), ax_6) \\
lcl'_7 &\equiv (\neg \text{ChiefActress}(a_2) \vee \text{Actress}(a_2), ax_1) & lcl'_8 &\equiv (\neg \text{ChiefActress}(a_2) \vee \neg \text{Man}(a_2), ax_2) \\
lcl'_9 &\equiv (\neg \text{Actress}(a_2) \vee \text{Person}(a_2), ax_3) & lcl'_{10} &\equiv (\neg \text{Actress}(a_2) \vee \neg \text{Actor}(a_2), ax_4) \\
lcl'_{11} &\equiv (\neg \text{Person}(a_2) \vee \text{Man}(a_2) \vee \text{Woman}(a_2), ax_5) & lcl'_{12} &\equiv (\text{Actor}(a_2), ax_7) \\
lcl'_{13} &\equiv (\text{ChiefActress}(a_1), ax_8) & lcl'_{14} &\equiv (\text{ChiefActress}(a_2), ax_9)
\end{aligned}$$

By applying the algorithm  $\text{Decompose}(\mathcal{G}_{bu}(\mathcal{L}(\mathcal{O})))$ , we can compute in turn that  $\Pi_{sc1} = \{lcl'_3, lcl'_5, lcl'_9, lcl'_{11}\}$  and  $\Pi_{sc2} = \{lcl'_2, lcl'_8\}$ , and compute the two maximal connected components of  $\mathcal{G}_{bu}(\mathcal{L}(\mathcal{O})) \setminus (\Pi_{sc1} \cup \Pi_{sc2})$  as  $\Pi_1 = \{lcl'_1, lcl'_4, lcl'_6, lcl'_{13}\}$  and  $\Pi_2 = \{lcl'_7, lcl'_{10}, lcl'_{12}, lcl'_{14}\}$ . It is not hard to see that the set  $\mathcal{M}_1$  of MISs of  $\text{Ax}(\Pi_1)$  is  $\{\{ax_1, ax_4, ax_6, ax_8\}\}$ , while the set  $\mathcal{M}_2$  of MISs of  $\text{Ax}(\Pi_2)$  is  $\{\{ax_1, ax_4, ax_7, ax_9\}\}$ . By Theorem 2, the set of all MISs of  $\mathcal{O}$  is  $\mathcal{M}_1 \cup \mathcal{M}_2 = \{\{ax_1, ax_4, ax_6, ax_8\}, \{ax_1, ax_4, ax_7, ax_9\}\}$ .

### 3.3 Using Approximate Compilation

The remaining problem of our proposed method is on the possible infiniteness of  $\mathcal{G}_{bu}(\mathcal{L}(\mathcal{O}))$ . We address this problem by using the same method given in [1]. That is, we introduce the notion of *convergent mapping function* for a labeled propositional program  $\Pi$  and compute a superset of  $\lambda(\mathcal{G}_{bu}(\mathcal{L}(\mathcal{O})))$  for  $\lambda$  a convergent mapping function for  $\mathcal{G}_{bu}(\mathcal{L}(\mathcal{O}))$ . A convergent mapping function  $\lambda$  for a labeled propositional program  $\Pi$  is defined as a mapping function from ground terms occurring in  $\Pi$  to constants occurring in  $\Pi$ , such that (i) for every functional term  $f_1(\dots f_n(a))$  (where  $a$  is a constant) occurring in  $\Pi$ ,  $\lambda(f_1(\dots f_n(a))) = \lambda(a)$ , and (ii) for every equational atom  $s \approx t$  occurring in  $\Pi$ ,  $\lambda(s) = \lambda(t)$ . The mapping function  $\lambda$  is further extended to a clause  $cl$  (resp. a labeled propositional program  $\Pi$ ) by defining  $\lambda(cl)$  (resp.  $\lambda(\Pi)$ ) as the result obtained from  $cl$  (resp.  $\Pi$ ) by replacing every ground term  $t$  occurring in it with  $\lambda(t)$ . The idea for introducing the convergent mapping function is illustrated the following figure.



As shown in the above figure (where arrows denote the sources of labeled clauses),  $\mathcal{G}_{bu}(\mathcal{L}(\mathcal{O}))$  is infinite even when  $\mathcal{O}$  consists of only three axioms. But after we map functional terms occurring in  $\mathcal{G}_{bu}(\mathcal{L}(\mathcal{O}))$  to constants and map a constant to another one if they occur in the same equational atom, we can obtain a small labeled propositional program  $\lambda(\mathcal{G}_{bu}(\mathcal{L}(\mathcal{O})))$  for  $\lambda$  a convergent mapping function for  $\mathcal{G}_{bu}(\mathcal{L}(\mathcal{O}))$ . In fact,  $\lambda(\mathcal{G}_{bu}(\mathcal{L}(\mathcal{O})))$  has polynomial number of labeled clauses wrt the size of  $\mathcal{O}$  because structural transformation guarantees that  $\mathcal{L}(\mathcal{O})$  has polynomial number of labeled clauses, and there are at most two distinct constants occurring in every labeled clause in  $\lambda(\mathcal{G}_{bu}(\mathcal{L}(\mathcal{O})))$ .

The following theorem shows that a set of sub-ontologies, from which all MISs of  $\mathcal{O}$  can be computed, are also computable from a superset of  $\lambda(\mathcal{G}_{bu}(\mathcal{L}(\mathcal{O})))$ .

**Theorem 3.** *Let  $\lambda$  be a convergent mapping function for  $\mathcal{G}_{bu}(\mathcal{L}(\mathcal{O}))$ . Let  $\Pi$  be a superset of  $\lambda(\mathcal{G}_{bu}(\mathcal{L}(\mathcal{O})))$ . Let  $\{\Pi\}_{1 \leq i \leq m}$  be returned by  $Decompose(\Pi)$ . Let  $\mathcal{M}_i$  be the set of MISs of  $Ax(\Pi_i)$  for  $1 \leq i \leq m$ . Then  $\bigcup_{i=1}^m \mathcal{M}_i$  is the set of all MISs of  $\mathcal{O}$ .*

*Proof.* Let  $\Pi_{sc1}$  and  $\Pi_{sc2}$  be the homonymous sets of labeled clauses computed in  $Decompose(\Pi)$ . Let  $\Pi'_{sc1} = \{(cl, ax) \in \mathcal{G}_{bu}(\mathcal{L}(\mathcal{O})) \mid (\lambda(cl), ax) \in \Pi_{sc1}\}$  and  $\Pi'_{sc2} = \{(cl, ax) \in \mathcal{G}_{bu}(\mathcal{L}(\mathcal{O})) \mid (\lambda(cl), ax) \in \Pi_{sc2}\}$ . It can be seen that, for all  $(cl, ax) \in \Pi'_{sc1}$ ,  $cl^+ \not\subseteq atoms(Cl(\mathcal{G}_{bu}(\mathcal{L}(\mathcal{O})) \setminus \Pi'_{sc1}))$ , and that for all  $(cl, ax) \in \Pi'_{sc2}$ ,  $cl^- \not\subseteq \bigcup_{(cl, ax) \in \mathcal{G}_{bu}(\mathcal{L}(\mathcal{O})) \setminus (\Pi'_{sc1} \cup \Pi'_{sc2})} cl^+$ . Let  $\{\Pi'_i\}_{1 \leq i \leq m'}$  be the

set of maximal connected components of  $\mathcal{G}_{bu}(\mathcal{L}(\mathcal{O})) \setminus (\Pi'_{sc1} \cup \Pi'_{sc2})$ . Let  $\mathcal{M}'_i$  be the set of MISs of  $\text{Ax}(\Pi'_i)$  for  $1 \leq i \leq m'$ . By Theorem 1,  $\bigcup_{i=1}^{m'} \mathcal{M}'_i$  is the set of all MISs of  $\mathcal{O}$ . Consider each  $1 \leq i \leq m'$ . Let  $\Pi''_i = \{(\lambda(cl), ax) \mid (cl, ax) \in \Pi'_i\}$ . It can be seen that  $\Pi''_i \subseteq \Pi \setminus (\Pi_{sc1} \cup \Pi_{sc2})$ . Since  $\Pi'_i$  is a connected component,  $\Pi''_i$  is a connected component of  $\Pi \setminus (\Pi_{sc1} \cup \Pi_{sc2})$ . Hence there exists  $\Pi_k$  ( $1 \leq k \leq m$ ) such that  $\Pi''_i \subseteq \Pi_k$ . It follows that  $\text{Ax}(\Pi'_i) = \text{Ax}(\Pi''_i) \subseteq \text{Ax}(\Pi_k)$ , then  $\mathcal{M}'_i \subseteq \mathcal{M}_k$ . Thus  $\bigcup_{i=1}^{m'} \mathcal{M}'_i \subseteq \bigcup_{i=1}^m \mathcal{M}_i$ . Since  $\mathcal{M}_i$  consists of only MISs of  $\mathcal{O}$ ,  $\bigcup_{i=1}^m \mathcal{M}_i$  is the set of all MISs of  $\mathcal{O}$ .  $\square$

To summarize, our proposed O-PP compilation process is to compile  $\mathcal{O}$  to a superset  $\Pi$  of  $\lambda(\mathcal{G}_{bu}(\mathcal{L}(\mathcal{O})))$  for  $\lambda$  a convergent mapping function for  $\mathcal{G}_{bu}(\mathcal{L}(\mathcal{O}))$ , while the inverse process of the O-PP compilation is to extract axiom parts from the result of `Decompose`( $\Pi$ ). To compute such  $\Pi$ , we adapt Algorithm 2 given in [1] by considering labels. Due to the space limitation, we do not provide the adapted algorithm here but refer the interested reader to [1] for more details.

## 4 Experimental Evaluation

We implemented a prototype system<sup>4</sup> that applies the Pellet [11] (version 2.0.1) debugging facility to compute MISs in every extracted sub-ontology. There is an optimization in our implementation. That is, before computing MISs in extracted sub-ontologies, the sub-ontologies having the same set of terminological axioms are combined into a bin and the consistency of the bin is checked. Only sub-ontologies not in any consistent bin are considered when computing MISs.

We used two groups of inconsistent ontologies. The test ontologies in the first group, including University, Chemical and MiniTambis, are modified from homonymous incoherent ontologies<sup>5</sup> by adding a concept assertion for every atomic concept. The test ontologies in the second group are modified from the University Benchmark (UOBM-Lite) [7] ontologies<sup>6</sup> by inserting a specified number of conflicts using the `Injector` tool described in [2], where a conflict is a set of axioms violating a functional role restriction or a disjointness constraint. By UOBM-Lite $_{n+m}$  we denote an UOBM-Lite ontology with assertional axioms of  $n$  universities and with  $m$  conflicts inserted. The characteristics of all test ontologies are given in Table 1. All experiments were conducted on a PC with Pentium Dual Core 2.60GHz CPU and 2GB RAM, running Windows XP, where the maximum Java heap size was set to (max) 1280MB for applying Pellet.

We compared our implemented system with Pellet [11] (version 2.0.1)<sup>7</sup> on computing MISs of all test ontologies. Both systems were set to use the glass box debugging method. Typical comparison results on execution time and some runtime statistics of our system are shown in Table 2. For two out of the three

<sup>4</sup> See <http://jfdw.limewebs.com/dbo-debug/> for more details on our system.

<sup>5</sup> <http://www.mindswap.org/ontologies/debugging/>

<sup>6</sup> <http://www.alphaworks.ibm.com/tech/semanticstk/>

<sup>7</sup> To the best of our knowledge, Pellet is the only existing OWL DL reasoner that provides debugging facilities.

**Table 1.** The characteristics of all test ontologies

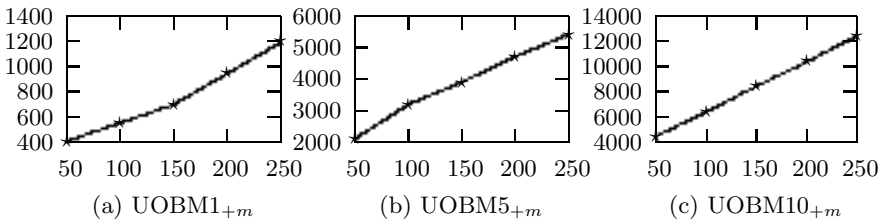
$\mathcal{O}$	Expressivity	#C	#R	#I	#Ax
University	<i>SOLF(D)</i>	30	12	34	74
Chemical	<i>ALCHF</i>	48	20	48	162
MiniTambis	<i>ALCF</i>	183	44	183	350
UOBM-Lite1 <sub>+50~+250</sub>	<i>SHIF(D)</i>	51	43	95,113–95,522	246,144–246,744
UOBM-Lite5 <sub>+50~+250</sub>				420,251–420,662	1,075,340–1,075,940
UOBM-Lite10 <sub>+50~+250</sub>				820,358–820,958	2,097,253–2,097,853

Note: “#C”, “#R”, “#I” and “#Ax” are respectively the numbers of atomic concepts, atomic roles, individuals and axioms in  $\mathcal{O}$ .

**Table 2.** Typical comparison results and some runtime statistics

$\mathcal{O}$	Pellet	Ours	Compile	Decompose	#MIS	#Sub	#A-S	#M-S
University	0:02:40	0:00:25	0:00:04	0:00:01	8	8	10	1
Chemical	>4:00:00	0:02:39	0:00:16	0:00:01	362	37	31	21
MiniTambis	>4:00:00	0:03:10	0:00:59	0:00:01	37	30	19	2
UOBM-Lite1 <sub>+50</sub>	>4:00:00	0:06:47	0:01:33	0:00:03	50	47	7	2
UOBM-Lite1 <sub>+250</sub>	>4:00:00	0:20:04	0:01:38	0:00:04	250	140	29	9
UOBM-Lite5 <sub>+50</sub>	>4:00:00	0:35:02	0:10:08	0:00:28	50	50	5	1
UOBM-Lite10 <sub>+50</sub>	out of mem	1:14:33	0:15:36	0:00:32	50	50	5	1

Note: “Pellet” (resp. “Ours”) is the total time Pellet (resp. our system) spends to compute all MISs of  $\mathcal{O}$ ; “Compile” (resp. “Decompose”) is the time spent in the O-PP compilation process (resp. the decomposition process); “#MIS” is the number of MISs of  $\mathcal{O}$ ; “#Sub” is the number of extracted sub-ontologies that are not in any consistent bin; “#A-S” (resp. “#M-S”) is the maximum number of axioms (resp. MISs) in every extracted sub-ontology that is not in any consistent bin.



**Fig. 1.** The total execution time (seconds) against increasing numbers of conflicts

ontologies in the first group, Pellet cannot compute all MISs in four hours, while our system computes all MISs in a few minutes. For all ontologies in the second group, Pellet cannot compute all MISs in four hours or runs out of memory when loading the test ontology, while our system works well too. We also used Pellet to debug the original incoherent ontologies in the first group and found that Pellet works well when there is no assertional axioms. These results show that Pellet is hard to debug an inconsistent ontology having hundreds of axioms and some

dozens of MISs where some axioms are assertional ones. In all our experiments, however, the Pellet debugging facility applied in our system only works on some extracted sub-ontology having at most 31 axioms and 21 MISs. This explains why our system works well and why it can significantly outperform Pellet.

We also tested our system against different numbers of inserted conflicts. As shown in Figure 1, our system scales well up to hundreds of conflicts when the number of conflicts increases.

## 5 Concluding Remarks

Debugging an inconsistent OWL DL ontology (i.e. computing all MISs of it) is an important problem in the Semantic Web, but currently lacks solutions that scale to large ontologies. For this situation, we proposed a solution to optimize the computation of all MISs and empirically showed its effectiveness in improving the scalability. The solution is based on a new compilation method from OWL DL to Propositional Logic (PL) and a new decomposition method on PL. Though our solution provides optimizations to existing OWL DL debugging facilities, the optimizations may not be effective in all possible cases. Hence we will research on approximate debugging methods in the future work.

## References

1. Du, J., Qi, G., Ji, Q.: Goal-directed module extraction for explaining OWL DL entailments. In: Bernstein, A., Karger, D.R., Heath, T., Feigenbaum, L., Maynard, D., Motta, E., Thirunarayan, K. (eds.) ISWC 2009. LNCS, vol. 5823, pp. 163–179. Springer, Heidelberg (2009)
2. Du, J., Shen, Y.: Computing minimum cost diagnoses to repair populated DL-based ontologies. In: Proc. of WWW 2008, pp. 265–274 (2008)
3. Fitting, M.: First-order Logic and Automated Theorem Proving, 2nd edn. Springer, Secaucus (1996)
4. Haase, P., van Harmelen, F., Huang, Z., Stuckenschmidt, H., Sure, Y.: A framework for handling inconsistency in changing ontologies. In: Gil, Y., Motta, E., Benjamins, V.R., Musen, M.A. (eds.) ISWC 2005. LNCS, vol. 3729, pp. 353–367. Springer, Heidelberg (2005)
5. Horrocks, I., Patel-Schneider, P.F., van Harmelen, F.: From *SHIQ* and RDF to OWL: the making of a web ontology language. *Journal of Web Semantics* 1(1), 7–26 (2003)
6. Kazakov, Y., Motik, B.: A resolution-based decision procedure for *SHOIQ*. *Journal of Automated Reasoning* 40(2-3), 89–116 (2008)
7. Ma, L., Yang, Y., Qiu, Z., Xie, G., Pan, Y., Liu, S.: Towards a complete OWL ontology benchmark. In: Sure, Y., Domingue, J. (eds.) ESWC 2006. LNCS, vol. 4011, pp. 125–139. Springer, Heidelberg (2006)
8. Parsia, B., Sirin, E., Kalyanpur, A.: Debugging OWL ontologies. In: Proc. of WWW 2005, pp. 633–640 (2005)
9. Reiter, R.: A theory of diagnosis from first principles. *Artificial Intelligence* 32(1), 57–95 (1987)

10. Schlobach, S., Cornet, R.: Non-standard reasoning services for the debugging of description logic terminologies. In: Proc. of IJCAI 2003, pp. 355–362 (2003)
11. Sirin, E., Parsia, B., Grau, B.C., Kalyanpur, A., Katz, Y.: Pellet: A practical OWL-DL reasoner. *Journal of Web Semantics* 5(2), 51–53 (2007)
12. Suntisrivaraporn, B., Qi, G., Ji, Q., Haase, P.: A modularization-based approach to finding all justifications for owl dl entailments. In: Domingue, J., Anutariya, C. (eds.) ASWC 2008. LNCS, vol. 5367, pp. 1–15. Springer, Heidelberg (2008)

# A Concept Hierarchy Based Ontology Mapping Approach

Ying Wang, Weiru Liu, and David Bell

School of Electronics, Electrical Engineering and Computer Science,  
Queen's University Belfast, Belfast, BT7 1NN, UK  
{`ywang14,w.liu,da.bell`}@qub.ac.uk

**Abstract.** Ontology mapping is one of the most important tasks for ontology interoperability and its main aim is to find semantic relationships between entities (i.e. concept, attribute, and relation) of two ontologies. However, most of the current methods only consider one to one (1:1) mappings. In this paper we propose a new approach (CHM: Concept Hierarchy based Mapping approach) which can find simple (1:1) mappings and complex (m:1 or 1:m) mappings simultaneously. First, we propose a new method to represent the concept names of entities. This method is based on the hierarchical structure of an ontology such that each concept name of entity in the ontology is included in a set. The parent-child relationship in the hierarchical structure of an ontology is then extended as a set-inclusion relationship between the sets for the parent and the child. Second, we compute the similarities between entities based on the new representation of entities in ontologies. Third, after generating the mapping candidates, we select the best mapping result for each source entity. We design a new algorithm based on the Apriori algorithm for selecting the mapping results. Finally, we obtain simple (1:1) and complex (m:1 or 1:m) mappings. Our experimental results and comparisons with related work indicate that utilizing this method in dealing with ontology mapping is a promising way to improve the overall mapping results.

## 1 Introduction

Research and development on ontology mapping (or matching) has attracted huge interests (e.g., [1,2,3,4,5,6]) and many mapping methods have been proposed. Comprehensive surveys on recent developments of ontology mapping can be found in [7,8].

Considerable efforts have been devoted to implement ontology mapping systems, especially one to one mappings. However, complex mappings (m:1, 1:m and m:n) are also pervasive and important in real world applications. In [7], an example was given to illustrate the importance of complex mappings in schema mapping research. We believe that the same issue exists in ontology mapping. Therefore, it is very important to find simple and complex mapping results in a natural way.

To address this problem in this paper, we first propose a new method to represent entities in ontologies. Traditionally, the concept names of entities are used



directly. This representation method does not consider the hidden relationships between concept names of entities, so it cannot reflect the complete meaning of the concept names of entities. When computing the similarities between entities based on this representation method, the result is hardly accurate. So it is significant to have a better method to represent entities. In this paper, we propose a new representation method for entities. For the multi-hierarchical structure of ontology, we view it as a concept hierarchy. For the example given in Figure 1(a), we observe that for each concept (in this paper, *concept*, *concept node* and *entity* represent the same thing) in this concept hierarchy, its complete meaning is described by a set of concept names. In other words, there is a kind of *semantic inclusion relationship* among these concepts. For instance, a branch from **CS**, **Courses** to **Graduate Courses** in Figure 1(a), **CS** means the department of computer science, **Courses** means the courses offered by the department of computer science and **Graduate Courses** means **Graduate Courses** is a kind of **Courses** and is offered by the department of computer science, i.e. **CS**, so the semantics of **Courses** can be completed by extending **Courses** to **{CS, Courses}**. Identically, we can extend the concept **Graduate Courses** to **{CS, Courses, Graduate Courses}**. Actually, a branch from one concept node to the root node indicates a complete meaning for this concept node. So for any concept name of entity  $C$  in an ontology, we can represent it by a new method as follows. First, we find the branch which has the concept  $C$ . Second, we collect those concepts along the path between  $C$  and the root node to form a set. We use this new set to represent entity  $C$ .

Once each entity is represented by a set of words, we compute the similarities between entities. In this paper, we separate the similarity values into two types: one is the similarities between entities which belong to one ontology, another is the similarities between entities which belong to two different ontologies. Here, we choose the *Linguistic-based matcher* (which uses domain specific thesauri to match words) and the *Structure-based matcher* (which uses concept-hierarchy theory) to compute similarities (we utilize Linguistic-based matcher because the performance of this matcher is good for similar or dissimilar words. Please refer to [9] for details).

As a result, we obtain a set  $S_1$  consisting of mapping candidates such that from each entity in ontology  $O_1$ , a similarity value is obtained for every entity in ontology  $O_2$ . Following this, we select the best mapping entity in  $O_2$  for each entity in  $O_1$  and these best mapping results constitute another set  $S_2$ . In  $S_2$ , we search all the mapping results to see if there exist multiple source entities in  $O_1$  that are mapped to the same target entity in  $O_2$ . If so, we apply a new algorithm based on Apriori algorithm [10] to decide how many source entities in  $O_1$  should be combined together to map onto the same entity in  $O_2$ . Our study shows that this method improves the matching results as illustrated in our experiments.

The rest of the paper is organized as follows. Section 2 describes the similarity measures used. Section 3 illustrates how to select final mapping results by using our new algorithm. Section 4 gives the background information about the

experiments and the results. Section 5 discusses related work and concludes the paper with discussions on future research.

## 2 Ontology Mapping

Ontology mapping can be done based on similarities, so we need to leverage the degree of the similarity between any two entities no matter these entities are in one ontology or from two ontologies. In this section, we describe our notion to measure the similarity between entities in detail. In this paper, we use *concept node* to denote an *entity* in ontology and we compute the similarity between concept nodes to indicate the similarity between entities. We compute the similarity of two concept nodes,  $e_i$  and  $e_j$ , denoted as  $Sim(e_i, e_j)$ :

$$Sim(e_i, e_j) = \begin{cases} \omega_{ls}sim_{ls}(e_i, e_j) + \omega_{ss}sim_{ss}(e_i, e_j) & e_i, e_j \in \text{same ontology} \\ sim_{ls}(e_i, e_j) & e_i, e_j \in \text{different ontologies} \end{cases} \quad (1)$$

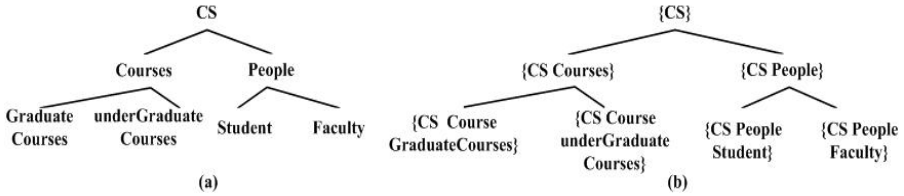
where  $\omega_{ls}$  and  $\omega_{ss}$  are two weight coefficients, reflecting the relative importance of the components. In our approach, we think both of the components are equally important, so we assign them both with coefficient 0.5.  $sim_{ls}(e_i, e_j)$  and  $sim_{ss}(e_i, e_j)$  denote the similarities obtained from the linguistic-based matcher and structure-based matcher respectively.

### 2.1 Extension for Concept Nodes

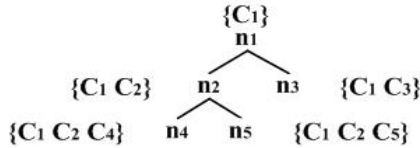
When using different methods to compute the similarity values between two names of entities in ontologies, such as *edit distance-based method* [11], or *Jaro-Winkler distance-based method* [12] etc, we discover that these methods are too simple to reflect the semantic relationship between those entities.

*Example 1.* Figure 1(a) provides a simple ontology which describes a department of computer science and shows its concept hierarchy. It is clear that if we only use one method (*edit distance-based method* or *Jaro-Winkler distance-based method*) to compute the similarity between those entities, such as “CS” and “People”, we cannot obtain good results because these two names of entities are not very similar directly but are related indirectly.

As shown in Figure 1(a), we found that the hierarchical structure is very similar to the *concept hierarchies* in multi-hierarchy association rules mining [10]. In this kind of mining, a concept hierarchy defines a sequence of mappings from a set of low-level concepts to higher-level, more general concepts. According to the concept hierarchies, “People” actually means “People in the department of Computer Science”, i.e. “People” and “CS” should be denoted as: {CS People} and {CS} separately. So we can denote all the concept names of entities in an ontology by a new approach in terms of the *inclusion relationship* between these concept names from the root node to leaf nodes and then Figure 1(a) can be changed to Figure 1(b).



**Fig. 1.** (a) An ontology which describes a department of computer science; (b) A new method to represent the ontology where we expand all the concepts



**Fig. 2.** An expanded ontology model where each node represents a concept

We now give precise similarity measures between entities. As stated above, concept names of entities have been expanded to concept sets, such as Figure 2, so we compute the similarity between any two sets by adopting a general method for computing similarities between composite words. We will introduce the method in the subsection: Calculating Similarities of Ontology Entities. So Equation(1) can be modified as:

$$Sim(E_i, E_j) = \begin{cases} \omega_{ls} sim_{ls}(E_i, E_j) + \omega_{ss} sim_{ss}(E_i, E_j) & E_i, E_j \in \text{same ontology} \\ sim_{ls}(E_i, E_j) & E_i, E_j \in \text{different ontologies} \end{cases} \quad (2)$$

where  $E_i, E_j$  are the concept sets for the concept nodes separately.

### 2.2 Linguistic-Based Matcher

We employ the *Linguistic-based matcher* as our similarity measure and in our paper the linguistic similarity between two concept nodes is denoted as  $sim(e_i, e_j)$ . *Linguistic-based matcher* uses common knowledge or domain specific thesauri to match words and this kind of matchers have been used in many papers [13,14].

The concept names of entities are usually composed of several words, so first we adopt Lin’s matcher to compute the similarity between two words and then we use another method to compute the similarity between concept names based on the revised version of Lin’s matcher.

**Lin’s Matcher:** Lin’s matcher is a kind of *Linguistic-based matcher*. In this paper, we use an electronic lexicon WordNet for calculating the similarity values between words. Lin in [15] proposed a probabilistic model which depends on corpus statistics to calculate the similarity values between words using the WordNet. This method is based on statistical analysis of corpora, so it considers

the probability of  $word_1$  ( $sense_1$ ) and  $word_2$  ( $sense_2$ ) and their most specific common subsumer  $lso(w_1, w_2)$  appearing in the general corpus. However, since the words in given ontologies are usually application specific, this general corpus statistics obtained using the WordNet can not reflect the real possibility of domain-specific words. To improve Lin's method, we propose to calculate a punishment coefficient according to the ideas in the path length method [16]. The path length method regards WordNet as a graph and measures the similarity between two concepts (words) by identifying the minimum number of edges linking the concepts. It provides a simple approach to calculating similarity values and does not suffer from the disadvantage that Lin's method does, so we integrate Lin's method and a punishment coefficient to calculate the similarity values between words. First, we outline Lin's approach. The main formulas in this method are as follows:  $sim_{Lin}(s_1, s_2) = \frac{2 \cdot \log(p(s_1, s_2))}{\log(p(s_1)) + \log(p(s_2))}$ ,  $p(s) = \frac{freq(s)}{N}$  and  $freq(s) = \sum_{n \in words(s)} count(n)$  where:  $p(s_1, s_2)$  is the probability that the same hypernym of sense  $s_1$  and sense  $s_2$  occurs,  $freq(s)$  denotes the word counts in sense  $s$ ,  $p(s)$  expresses the probability that sense  $s$  occurs in some synset and  $N$  is the total number of words in WordNet.

The punishment coefficient which is based on the theory of path length of WordNet is denoted as:  $\frac{1}{2}\alpha^l$ . Its meaning is explained as follows:  $\alpha$  is a constant between 0 and 1 and is used to adjust the decrease of the degree of similarity between two senses when the path length between them is deepened and  $l$  expresses the longest distance either sense  $s_1$  or sense  $s_2$  passes by in a hierarchical hypernym structure. Because sense  $s_1$  and sense  $s_2$  occupy one of the common branches, this value has to be halved.

Therefore in our method, the similarity value calculated by Lin's method is adjusted with this coefficient to reflect more accurate degree between two senses  $s_1$  and  $s_2$ . The revised calculation is:

$$sim_{new}(s_1, s_2) = \frac{2 \cdot \log(p(s_1, s_2))}{\log(p(s_1)) + \log(p(s_2))} \bullet \frac{1}{2}\alpha^l \quad (3)$$

Word  $w_1$  and word  $w_2$  may have many senses, we use  $s(w_1)$  and  $s(w_2)$  to denote the sets of senses for word  $w_1$  and word  $w_2$  respectively as  $s(w_1) = \{s_{1i} \mid i = 1, 2, \dots, m\}$ ,  $s(w_2) = \{s_{2j} \mid j = 1, 2, \dots, n\}$ , where the numbers of senses that word  $w_1$  and word  $w_2$  contain are  $m$  and  $n$ . We then choose the maximum similarity value between two senses from the two sets of senses for words  $w_1$  and  $w_2$ , so the similarity between words is:  $sim(w_1, w_2) = \max(sim_{new}(s_{1i}, s_{2j}))$ ,  $1 \leq i \leq m, 1 \leq j \leq n$

**Calculating Similarities of Ontology Entities:** We compute similarities between names of ontology entities based on the word similarities obtained from the two matchers separately. The names of ontology entities are composed of several words, so we split a phrase (name of entity) and put the individual words into a set and then we deal with these words as follows: first, we calculate similarities of every pair of words within both sets by using one of the matchers (Linguistic-based matcher or Structure-base matcher). After that, for each word

in one set, compute similarity values between this word and every word from the other set and then pick out the largest similarity value. Finally attach this value to the word. Repeat this step until all of the words in the two sets have their own values. Finally, we compute the final degree of similarity of names using the sum of similarity values of all words from two sets divided by the total counts of all words.

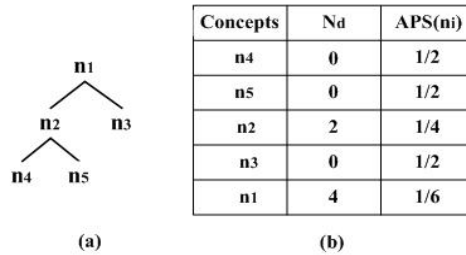
### 2.3 Structure-Based Matcher

Ontology can be regarded as a model of multi-hierarchy, so in terms of the structure we propose a *Structure-based Matcher* which determines the similarity between two nodes (entities) based on the number of children nodes. We first introduce the method.

An ontology is usually designed in such a way that its topology and structure reflects the information contained within and between the concepts. In [17], Schickel-Zuber and Faltings propose the computation of the *a-priori score* of concept  $c$ ,  $APS(c)$ , which captures this information. Equation (4) gives the definition of the a-priori score of concept  $c$  with  $n$  descendants as:

$$APS(c) = \frac{1}{n + 2} \quad (4)$$

To illustrate the computation of the a-priori score, consider the simple example shown in Figure 3 where  $n_i$  represents the concept node in the ontology and  $N_d$  is the number of descendants for each concept node  $n_i$ . First, the number of descendants of each concept is computed. Then, Equation (8) is applied to compute the a-priori score of each concept in the ontology.



**Fig. 3.** (a) An ontology model where each node represents a concept; (b) The a-priori scores for those concepts

It is very easy to find that the concept becomes more generalized as we travel up the ontology, and the a-priori score decreases due to the increasing number of descendants. That is the a-priori score reflects the information of each concept node, i.e., the higher score the concept has, the more information the concept expresses. So it is possible to estimate the similarity between two concept nodes by finding the overlapping part of information between two concepts.

After obtaining the a-priori score for each concept node, we use the following definition to calculate the similarity as the structure-based matcher.

Given two a-priori scores  $APS(n_i)$  and  $APS(n_j)$  for two concept nodes  $n_i$  and  $n_j$  respectively, the similarity between  $n_i$  and  $n_j$  is defined as [17]:

$$sim_{ss}(n_i, n_j) = \frac{\min(APS(n_i), APS(n_j))}{\max(APS(n_i), APS(n_j))} \quad (5)$$

*Example 2.* From Figure 3(b), we can get the  $APS(n_i)$  value for each node  $n_i$  and then we can compute the similarity between any two nodes. For instance,  $sim_{ss}(n_1, n_4) = \frac{1/6}{1/2} = 1/3$ .

### 3 Selection of the Best Mapping Results

For each entity  $e_i$  in  $O_1$ , we apply the linguistic-based matcher for computing the similarities between this entity and every member of  $O_2$  and find the best mapping for this entity. Let  $S$  denote the set that contains the best mapping candidate in  $O_2$  for every entity in  $O_1$ . In  $S$ , there may exist complex mapping results, i.e. several entities in  $O_1$  map to the same entity in  $O_2$ . Our task is to decide where several entities in  $O_1$  should be mapped to the same entity in  $O_2$ .

DCM framework [18] is a schema matching system and it is focused on dealing with matching multiple schemas together. In this framework, there is a APRIORICORRMING algorithm for discovering correlated items. In [18], correlated items are defined as the mapping results. This algorithm is to find all the correlated items with size  $l+1$  based on the ones with size  $l$  in multiple schemas. It first finds all correlated two items and then recursively constructs correlated  $l+1$  items from correlated  $l$  items. In this paper, our aim is to make sure if several entities in  $O_1$  should be combined together to map to the same entity in  $O_2$ , so we regard the entities in  $O_1$  as the items and attempt to find if they are correlated. We try to obtain the most correlated items directly, but APRIORICORRMING algorithm is not suitable for our objective, so we propose an improved algorithm named REVISEDAPRIORIMING based on APRIORICORRMING algorithm.

First, for each entity of  $O_2$  in set  $S$ , we collect its mapping entities of  $O_1$  and put them into set  $Z$  such that  $Z \subseteq O_1$ . And then input  $Z$  and use the REVISEDAPRIORIMING to find if these entities should be combined together to map one entity in  $O_2$ . As shown in Algorithm 1, first we find the incorelate entities in  $V$  based on the similarities between them and store them into  $X$  (Line 4-8). Next, for each item in  $X$ , we have to construct different entities groups in which two entities of one item in  $X$  can not happen together (Line 9-16). When this algorithm is complete, we obtain a set  $V$  that stores the entity groups. Each entity group is a different combination of correlated entities. We search the set  $V$  to find the largest entity groups (in terms of cardinality). Since there may exist more than one such group, i.e. the number of entities in these groups are the same, we select one such group by using the formula below:

$$G_e = \arg \max_{k=1}^n ( \sum_{i=1}^n sim(e_i, e_j) ), e_i \in O_1, e_j \in O_2 \quad (6)$$

where  $G_e$  denotes the entity group which stores the combined entities,  $l$  is the number of entity groups in  $V$  and  $n$  is the number of entities in each group.

---

**Algorithm 1.** REVISED APRIORI MINING
 

---

**Input:** Input entities in  $\mathcal{O}_1$ :  $Z = \{e_1, e_2, \dots, e_n\}$ , Threshold  $T$

**Output:** Combined entity groups  $V = \{V_1, V_2, \dots, V_m\}$

```

1:  $X \leftarrow \emptyset$ 
2: Create two queues  $A \leftarrow \emptyset, V \leftarrow \emptyset$ 
3:  $V = V \cup \{Z\}$ 
4: for all  $e_i, e_j \in Z, i \neq j$  do
5:   if  $sim(e_i, e_j) < T$  then
6:      $X \leftarrow X \cup \{e_i, e_j\}$ 
7:   end if
8: end for
9: for each item  $\{e_i, e_j\} \in X$  do
10:   $A = V, V = \emptyset$ 
11:  for each set  $V_s$  in  $A$  do
12:     $A = A \setminus \{V_s\}$ 
13:    Remove  $e_i$  and  $e_j$  respectively from  $V_s$ , then  $V_s$  is changed into two different sets  $V_p$  and  $V_q$ 
14:     $\hat{V} = V \cup \{V_p\}, V = V \cup \{V_q\}$ 
15:  end for
16: end for
17: return  $V$ 

```

---

## 4 Experiments

### 4.1 Dataset

We have implemented our approach in Java and now we present the experimental results that demonstrate the performance of our methods using the OAEI 2007 Benchmark Tests. In our experiments, we only focus on classes and properties in ontologies.

Generally, almost all the benchmark tests in OAEI 2007 describe Bibliographic references except Test 102 which is about wine and it is totally irrelevant to other data. We choose twenty-five test data for testing. All of these twenty-five test data can be divided into four groups [19] in terms of their characteristics: Test 101-104, Test 201-210, Test 221-247 and Test 301-304. A brief description is given below.

- **Test 101-104:** These tests contain classes and properties with either exactly the same or totally different names.
- **Test 201-210:** The tests in this group change some linguistic features compared to Test 101-104. For example, some of the ontologies in this group have no comments or names, names of some ontology have been replaced with synonyms.
- **Test 221-247:** The structures of the ontologies have been changed but the linguistic features have been maintained.
- **Test 301-304:** Four real-life ontologies about BibTeX.

In our evaluation, we take **Test 101** as the reference ontology. All the other ontologies are compared with **Test 101**.

## 4.2 Comparison of Experimental Results

We now compare the outputs from our system (denoted as *CHM*) to the results obtained from the best performing system: *ASMOV*, average performing systems: *DSSim*, *OntoDNA* and worse performing system: *TaxoMap* which were from the 2007 Ontology Alignment Contest<sup>1</sup>. There are fifty tests totally. However, in some of ontologies, the names of entities are represented in scramble or in French, so the similarities between the names of entities can not be computed by our linguistic-based matcher. We ignore the comparisons of these ontologies. The details of experimental results are given in Table 1. In Table 1,  $p$  for precision,  $r$  for recall,  $f$  for f-measure,  $Best$  and  $Worst$  denote the values  $\frac{f(SIM)}{f(Best\ or\ Worst)}$  between *CHM* and *Best* system or *Worst* system in one row. If the value equals to 1, it means these systems obtain the same results. If the value is smaller than 1, it means *CHM* presents worse than other systems. Otherwise, it shows *CHM* is better than others.

**Table 1.** Comparison of experiment results

Groups	Datasets	CHM			ASMOV			DSSim			TaxoMap			OntoDNA			Best	Worst
		p	r	f	p	r	f	p	r	f	p	r	f	p	r	f	f	f
Test 101-104	101	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	1	1
	103	100	100	100	100	100	100	100	100	100	100	34	51	94	100	97	1	1.96
	104	100	100	100	100	100	100	100	100	100	100	34	51	94	100	97	1	1.96
Test 201-210	203	100	100	100	100	100	100	100	100	100	NaN	0.00	NaN	94	100	97	1	$\infty$
	204	86	84	85	100	100	100	96	91	93	92	24	38	93	84	88	0.85	2.24
	205	47	44	46	100	100	100	94	33	49	77	10	18	57	12	20	0.46	2.56
	208	86	83	85	100	100	100	95	90	92	NaN	0	NaN	93	84	88	0.85	$\infty$
	209	49	41	45	92	90	91	91	32	47	NaN	0	NaN	57	12	20	0.49	$\infty$
Test 221-247	221	82	82	82	100	100	100	100	100	100	100	34	51	93	76	83	0.82	1.61
	222	89	92	91	100	100	100	100	100	100	100	31	47	94	100	97	0.91	1.94
	224	100	100	100	100	100	100	100	100	100	100	34	51	94	100	97	1	1.96
	225	100	100	100	100	100	100	100	100	100	100	34	51	94	100	97	1	1.96
	228	100	100	100	100	100	100	100	100	100	100	100	100	53	27	36	1	2.78
	230	73	90	81	99	100	99	97	100	98	100	35	52	91	100	95	0.82	1.56
	231	100	100	100	100	100	100	100	100	100	100	34	51	94	100	97	1	1.96
	232	82	82	82	100	100	100	100	100	100	100	34	51	93	76	84	0.82	1.61
	233	52	52	52	100	100	100	100	100	100	100	100	100	53	27	32	0.52	1.63
	236	100	100	100	100	100	100	100	100	100	100	100	100	53	27	32	1	3.13
	237	93	97	95	100	100	100	100	100	100	100	31	47	94	100	97	0.95	2.02
	239	88	100	94	97	100	98	97	100	98	100	100	100	50	31	38	0.94	2.47
	241	58	58	58	100	100	100	100	100	100	100	100	100	53	27	32	0.58	1.81
246	88	100	94	97	100	98	97	100	98	100	100	100	50	31	38	0.94	2.47	
Test 301-304	301	43	45	44	93	82	87	82	30	44	100	21	35	88	69	77	0.51	1.26
	302	34	53	42	68	58	63	85	60	70	100	21	35	90	40	55	0.6	1.2
	304	51	49	50	95	96	95	96	92	94	93	34	50	92	88	90	0.53	1

Overall, we believe that the experimental results of our system are good. Although on individual pair of ontologies, our results are less ideal than the *ASMOV* system and *DSSim*, however, our results are better than *TaxoMap* system and *OntoDNA* system on most pairs of matching. The performances of these three different approaches, i.e., *ASMOV*, *DSSim* and our system *CHM* are good for almost the whole data set from Test 101 to Test 246, but our

<sup>1</sup> <http://oaei.ontologymatching.org/2007/results/>



system does not perform well for Test 205, Test 209, Test 233 and Test 241. The performance of all these five systems are not very good for the data set from Test 301 to Test 304. Below we analyze the reasons for this.

### One-to-one Mapping Results

- More effective results:
  - The two ontologies to be matched contain classes and properties with exactly the same names and structures, so every system that deploys the computation of similarities of names of entities can get good results, for instance, Test 101 vs 103 and vs 104.
  - Most of the results of Test 221-247 are good because the linguistic features have been maintained. However, the structures of them have been changed, so the performance of our system has been affected.
- Less effective results:
  - Test 201-210 describe the same kind of information as other ontologies, i.e. publications, however, the class names in them are very different from those in the reference ontology Test 101, especially Test 205 and 209, so our system does not obtain good results.
  - Our method is based on the hierarchical structure of an ontology, but for Test 233 and Test 241, these two ontologies have only one layer. When computing the similarity between two concepts in Test 233 and Test 101, such as **MastersThesis** in Test 233 and **MastersThesis** in Test 101. First, our method extends **MastersThesis**. Test 233 only has one layer, so **MastersThesis** can not be changed. Test 101 has three layers, so **MastersThesis** is extended to **{MastersThesis, Academic, Reference}**. The similarity value is reduced and does not reflect the true similarity between these two concepts.

**Table 2.** Comparison of complex mapping results

Group	Datasets	CHM		
		p	r	f
Test 301-304	301	55	55	55
	302	71	42	53
	304	33	50	40

**Complex Mapping Results** In Test 301-304, there exists inclusion relationships between entities, for example, **Collection**<**Book**. Several source entities have the inclusion relationships to one target entity separately, so we take this mapping as complex (m:1) mapping.

Tests 301-304 are real-life BibTeX ontologies which also include different words compared to Test 101 describing publications so the results are similar to Test 205, so we do not get good similarity results from this data set. However we still find some complex mappings (m:1) by using our algorithm to discover the best mapping results, such as for Test 302 vs Test 101, we get **{Collection, Monograph, Book}** mapping to **Book**.

## 5 Related Work and Conclusion

**Related Work:** Ontology matching is important and has received significant attention. However, existing matching methods mostly focus on simple (1:1) mappings. Here we present some related work on complex matching.

RiMOM [20] is a general ontology mapping system based on Bayesian decision theory and the approach divides the process of discovering complex mapping (m:1 mapping) into two steps: mapping entities discovery and mapping expression discovery. We mainly focus on mapping entities discovery. In mapping entities discovery, the system aims at finding whether there are multiple source entities mapped onto one target entity. If there are multiple source entities that are mapped onto one target entity, then it will combine those source entities. It does not consider any hidden relationship between those source entities, so it may cause wrong complex mappings.

PBM [21] and BMO [22] are two approaches to focusing on complex mappings so far. PBM is a method for partition-based block matching that is practically applicable to large class hierarchies. It first partitions the two large class hierarchies into blocks separately and then constructs the mappings between these blocks. It does not consider the relationship between the classes which belong to two large hierarchies respectively. It partitions these two hierarchies separately instead of partitioning one hierarchy by referring another one. This may produce wrong mappings. It is not always reasonable to partition a hierarchy itself into blocks, so this is one of the reason that it can not obtain very good results.

BMO tries to utilize a partition method to handle complex mappings just like PBM. But it is different on the process of partition. BMO puts the two ontologies together for partitioning and then obtains the block mapping results directly. It considers the correspondence between different ontologies and avoids too much manual intervention. BMO is a method that implements complex matching, but it failed to consider many issues, such as if it finds multiple complex mapping results, then what method should be used to combine them, etc.

**Conclusion:** In this paper, we proposed a new representation for concepts in ontology and then utilized two matchers (*Linguistic-based matcher* and *Structure-based matcher*) to deal with ontology mapping. In the *Linguistic-based matcher*, we improved Lin's method which computes similarity values between words. In the *Structure-based matcher*, we adopted the structure of ontology to calculate the similarity between two entities. Following this, we investigated how we can obtain reasonable ontology mapping results. We apply our new algorithm to search complex ontology mapping (m:1 mapping) from a set of ontologies used for ontology mapping competitions. The experimental results demonstrated that it is feasible for dealing with ontology mapping.

## References

1. Ehrig, M., Sure, Y.: Ontology mapping - an integrated approach. In: Bussler, C.J., Davies, J., Fensel, D., Studer, R. (eds.) ESWS 2004. LNCS, vol. 3053, pp. 76–91. Springer, Heidelberg (2004)

2. Ehrig, M., Staab, S.: Qom - quick ontology mapping. In: McIlraith, S.A., Plexousakis, D., van Harmelen, F. (eds.) ISWC 2004. LNCS, vol. 3298, pp. 683–697. Springer, Heidelberg (2004)
3. Noy, N.F., Musen, M.A.: Prompt: Algorithm and tool for automated ontology merging and alignment. In: Proceedings of the 17th National Conference on Artificial Intelligence and 12th Conference on Innovative Applications of Artificial Intelligence (AAAI/IAAI 2000), pp. 450–455 (2000)
4. Noy, N.F., Musen, M.A.: Anchor-prompt: Using non-local context for semantic matching. In: Workshop on Ontologies and Information Sharing at the 17th International Joint Conference on Artificial Intelligence, IJCAI 2001 (2001)
5. Kalfoglou, Y., Schorlemmer, W.M.: Information-flow-based ontology mapping. In: Meersman, R., Tari, Z., et al. (eds.) CoopIS 2002, DOA 2002, and ODBASE 2002. LNCS, vol. 2519, pp. 1132–1151. Springer, Heidelberg (2002)
6. Su, X., Gulla, J.A.: Semantic enrichment for ontology mapping. In: Meziane, F., Métais, E. (eds.) NLDB 2004. LNCS, vol. 3136, pp. 217–228. Springer, Heidelberg (2004)
7. Rahm, E., Bernstein, P.A.: A survey of approaches to automatic schema matching. *Journal of VLDB* 10(4), 334–350 (2001)
8. Shvaiko, P., Euzenat, J.: A survey of schema-based matching approaches. *Journal of Data Semantics* 4, 146–171 (2005)
9. Wang, Y., Liu, W., Bell, D.: Combining uncertain outputs from multiple ontology matchers. In: Prade, H., Subrahmanian, V.S. (eds.) SUM 2007. LNCS (LNAI), vol. 4772, pp. 201–214. Springer, Heidelberg (2007)
10. Han, J., Kamber, M.: *Data Mining: Concepts and Techniques* (2000)
11. Navarro, G.: A guided tour to approximate string matching. *ACM Computing Surveys* 33(1), 31–88 (2001)
12. Winkler, W.E.: The state of record linkage and current research problems. In: Proceedings of the Survey Methods Section Statistical Society of Canada, pp. 73–80 (1999)
13. Tang, J., Liang, B., Li, J.: Multiple strategies detection in ontology mapping. In: Proceedings of the 14th International Conference on World Wide Web (WWW 2005) (Special interest tracks and posters), pp. 1040–1041 (2005)
14. Madhavan, J., Bernstein, P.A., Rahm, E.: Generic schema matching with cupid. In: Proceedings of 27th International Conference on Very Large Data Bases (VLDB 2001), pp. 49–58 (2001)
15. Lin, D.: An information-theoretic definition of similarity. In: Proceedings of the 15th International Conference on Machine Learning (ICML 1998), pp. 296–304 (1998)
16. Resnik, P.: Using information content to evaluate semantic similarity in a taxonomy. In: Proceedings of 14th International Joint Conference for Artificial Intelligence (IJCAI 1995), pp. 448–453 (1995)
17. Schickel-Zuber, V., Faltings, B.: OSS: A semantic similarity function based on hierarchical ontologies. In: Proceedings of the 20th International Joint Conference on Artificial Intelligence (IJCAI 2007), pp. 551–556 (2007)
18. He, B., Chang, K.C.: Automatic complex schema matching across Web query interfaces: A correlation mining approach. *ACM Transactions on Database Systems* 31(1), 346–395 (2006)
19. Qu, Y., Hu, W., Cheng, G.: Constructing virtual documents for ontology matching. In: Proceedings of 15th World Wide Web Conference (WWW 2006), pp. 23–31 (2006)

20. Tang, J., Li, J., Liang, B., Huang, X., Li, Y., Wang, K.: Using Bayesian decision for ontology mapping. *Journal of Web Semantics* 4(4), 243–262 (2006)
21. Hu, W., Zhao, Y., Qu, Y.: Partition-based block matching of large class hierarchies. In: Mizoguchi, R., Shi, Z.-Z., Giunchiglia, F. (eds.) *ASWC 2006*. LNCS, vol. 4185, pp. 72–83. Springer, Heidelberg (2006)
22. Hu, W., Qu, Y.: Block Matching for Ontologies. In: Cruz, I., Decker, S., Allemang, D., Preist, C., Schwabe, D., Mika, P., Uschold, M., Aroyo, L.M. (eds.) *ISWC 2006*. LNCS, vol. 4273, pp. 300–313. Springer, Heidelberg (2006)

# Composing Cardinal Direction Relations Basing on Interval Algebra

Juan Chen, Haiyang Jia, Dayou Liu, Changhai Zhang

College of Computer Science and Technology, Jilin University, Changchun 130012, China  
Key Laboratory of Symbolic Computation and Knowledge Engineering of Ministry of  
Education, Jilin University, Changchun 130012, China  
{chenjuan, jiahy, dyliu, changhai}@jlu.edu.cn

**Abstract.** Direction relations between extended spatial objects are important commonsense knowledge. Skiadopoulos proposed a formal model for representing direction relations between compound regions (the finite union of simple regions), known as SK-model. It perhaps is currently one of most cognitive plausible models for qualitative direction information, and has attracted interests from artificial intelligence and geographic information system. Originating from Allen first using composition table to process time interval constraints; composing has become the key technique in qualitative spatial reasoning to check the consistency. Due to the massive number of basic directions in SK-model, its composition becomes extraordinary complex. This paper proposed a novel algorithm for the composition. Basing the concepts of smallest rectangular directions and its original directions, it transforms the composition of basic cardinal direction relations into the composition of interval relations corresponding to Allen's interval algebra. Comparing with existing methods, this algorithm has quite good dimensional extendibility, that is, it can be easily transferred to the tridimensional space with a few modifications.

**Keywords:** Cardinal direction relation, interval algebra, composing, qualitative spatial reasoning.

## 1 Introduction

Spatial representation and reasoning plays an essential role in human activities. Although the quantitative approaches can provide the most precise information, the numerical information is often not necessary or unavailable at the human level. Qualitative approach for spatial reasoning, known as qualitative spatial reasoning (QSR) [1], becomes a promising way to process spatial information at this level and has prevailed in artificial intelligence (AI) and geographical information systems (GIS) communities for over three decades. Consequently dozens of formalisms of spatial relations have been proposed *e.g.*, topology [2]-[3], directions [4]-[10] and combined [11]-[15].

This paper concentrates on the problem of composing cardinal directions. We start from the model of Skiadopoulos [8], [9] (SK-model) which is the currently one of most cognitive plausible models for qualitative reasoning with cardinal directions.



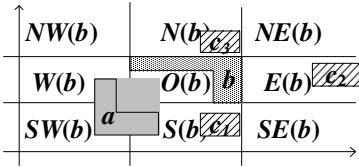


Fig. 1. Pictorial examples of basic cardinal direction

### 3. Interval Algebra

Interval algebra (IA) was proposed by Allen [16] for temporal reasoning. According to the different order relations between the temporal intervals' endpoints, it defines 13 basic interval relations, which constitutes a JEPD (Jointly Exhaustive and Pairwise Disjoint) list of all possible relations which can hold between temporal intervals, denoted by  $A_{int} = \{p, pi, m, mi, e, d, di, s, si, f, fi, o, oi\}$  as shown in Table 1. Each element of the power set  $2^{A_{int}}$  can be seen as the union of basic relations, which is used to describe the indeterminate information. e.g.  $\{e, m\}$  means the possible relations between the intervals are equals or meets.

Table 1. basic interval relations

relation	symbol	inverse	meaning	relation	symbol	inverse	meaning
precedes	$p$	$pi$		during	$d$	$di$	
meets	$m$	$mi$		finishes	$f$	$fi$	
overlaps	$o$	$oi$		equal	$e$	$e$	
starts	$s$	$si$					

When dealing with planar or cubic space, IA can be extended to rectangle algebra [17] and block algebra [18] which describe the spatial configurations by listing the relation for each coordinate separately, while it can only describe rectangles or blocks whose sides are parallel to the orthogonal basis. Here we only consider the Euclidean space  $\mathbb{R}^2$ , the relation between the given rectangles must be in the rectangle relations denoted by  $A_{rec} = \{(r_x, r_y) | r_x, r_y \in A_{int}\}$  [17]. As shown in Fig. 2, the relation between rectangle  $a$  and  $b$  is the pair  $(o, p)$ .

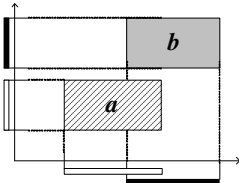


Fig. 2. picture of basic rectangle relation  $(o, p)$

### 4. Correlations between cardinal direction relations and interval algebra

According to the definitions of atomic cardinal direction relation, the projections of arbitrary region on each axis must satisfy a certain interval relation, as Table 2 shows.

**Table 2.** six groups of interval relations

Interval relation	Projection relation $\pi_{x,y}$
$\{p, m\}$	$sup_{\pi}(a) \leq inf_{\pi}(b)$
$\{o, fi\}$	$inf_{\pi}(a) < inf_{\pi}(b) < sup_{\pi}(a) \leq sup_{\pi}(b)$
$\{e, d, s, f\}$	$inf_{\pi}(b) \leq inf_{\pi}(a) \quad sup_{\pi}(a) \leq sup_{\pi}(b)$
$\{pi, mi\}$	$sup_{\pi}(b) \leq inf_{\pi}(a)$
$\{oi, si\}$	$inf_{\pi}(b) < inf_{\pi}(a) < sup_{\pi}(b) \leq sup_{\pi}(a)$
$\{di\}$	$inf_{\pi}(a) < inf_{\pi}(b) \quad sup_{\pi}(b) < sup_{\pi}(a)$

**Definition 1.** Basic direction  $r$  is rectangular iff there exists two rational rectangles  $a$  and  $b$  such that  $dir(a, b)=r$  is satisfied; otherwise, it is non-rectangular.

**Example 1.** all atomic directions are rectangular.  $N:NW$  is rectangular while  $NW:N:W$  is not.

Let  $r$  be a basic rectangular direction,  $REC(r)=REC_x(r) \times REC_y(r)$  is the corresponding rectangle relation of  $r$  according to Table 2. For example, if  $r=N$ , then  $REC(r)=\{e, d, s, f\} \times \{pi, mi\}$ . Contrariwise, for a rectangle relation  $T$ , its corresponding rectangular direction is  $CD(T)= \{CD(t) \mid t \in T\}$ . e.g.,  $T=\{s, o\} \times \{f\} \times \{o\}$ ,  $CD(T)= CD((s, f)) CD(o, f)=\{O, W:O\}$ .

**Definition 2.** Basic direction  $r_1=r_{11}:\dots:r_{1m}$  includes basic relation  $r_2=r_{21}:\dots:r_{2n}$  iff  $\{r_2, \dots, r_{2n}\} \subseteq \{r_{11}, \dots, r_{1m}\}$ , denoted by  $r_2 \subseteq r_1$ .

**Example 2.**  $N:O$  (properly) includes atomic relations  $N$  and  $O$ .

**Definition 3:** Basic direction  $r_1$  is the smallest rectangular direction (SRD) of  $r_2$ , denoted by  $SRD(r_2)$ , iff it is rectangular and the smallest direction includes  $r_2$ .

**Definition 4:** Basic directions whose smallest rectangular direction is  $r$  are called the original directions of  $r$ , denoted by  $ORG(r)$ .  $ORG(r)$  might be a single basic direction or a set of basic directions.

**Example 3.**  $ORG(W)=\{W\}$ ,  $ORG(NW:N:W:O)=\{N:W:O, NW:W:O, NW:N:O, NW:N:W, NW:N:W:O, NW:O, N:W\}$  over  $REG^*$ , when restricted to simple regions,  $NW:O$  and  $N:W$  can not be satisfied.

Based on above definitions, the set of basic directions can be divided into  $6^2=36$  equivalence classes, as shown in Table 3. Each element in each class is equal in its  $SRD$ , while each basic rectangle relation corresponds to a rectangular direction which is the  $SRD$  of a set of basic directions. Thus we can establish the mapping between basic cardinal direction and rectangle relations.

**Table 3.** Rectangle relations of rectangular direction

$REC_y$ ( $r$ ) $REC_x(r)$	$\{p, m\}$	$\{o, fi\}$	$\{e, d, s, f\}$	$\{si, oi\}$	$\{pi, mi\}$	$\{di\}$
$\{p, m\}$	$SW$	$SW:W$	$W$	$NW:W$	$NW$	$SW:W:NW$



$\{o, fi\}$	$SW:S$	$SW:S:W:O$	$W:O$	$NW:N:W:O$	$NW:N$	$NW:N:W:O:SW:S$
$\{e, d, s, f\}$	$S$	$S:O$	$O$	$N:O$	$N$	$N:O:S$
$\{si, oi\}$	$S:SE$	$S:SE:O:E$	$O:E$	$N:NE:O:E$	$N:NE$	$N:NE:O:E:S:SE$
$\{pi, mi\}$	$SE$	$E:SE$	$S:SE$	$NE:E$	$NE$	$NE:E:SE$
$\{di\}$	$SW:S:SE$	$W:O:E:SW:S:SE$	$W:O:E$	$NW:N:NE:W:O:E$	$NW:N:NE$	$NW:N:NE:W:O:E:SW:S:SE$

## 5. Composing

Originating from Allen [16] first using composition table to process time interval constraints; the notion of composition plays a very important role in qualitative spatial reasoning. It has become the key technique in QSR to check the consistency.

**Definition 5:** Let  $r_1$  and  $r_2$  be two basic directions, their composition, denoted by  $r_1 \circ r_2$  is a set of basic directions satisfying: for each  $t \in r_1 \circ r_2$ , there exist region  $a, b, c \in REG^*$  such that  $dir(a, b)=r_1 \quad dir(b, c)=r_2 \quad dir(a, c)=t$  is satisfied.

**Theorem 1:** For arbitrary basic directions  $r_1, r_2 \in D^*$ ,  $r_1 \circ r_2 = r_1 \circ SRD(r_2)$  holds.

**Proof:** For  $\forall t \in r_1 \circ r_2$ , there exist  $a, b, c \in REG^*$ , such that

$$dir(a, b)=r_1 \quad dir(b, c)=r_2 \quad dir(a, c)=t.$$

Then according to the definitions of cardinal directions and SRD,

$$dir(a, MBR(b))=r_1 \quad dir(MBR(b), MBR(c))=SRD(r_2) \quad dir(a, MBR(c))=t$$

So,  $t \in r_1 \circ SRD(r_2)$  holds.

Conversely,  $\forall t \in r_1 \circ SRD(r_2)$ , there exists region  $a, b, c \in REG^*$ , such that

$$dir(a, b)=r_1 \quad dir(b, c)=SRD(r_2) \quad dir(a, c)=t$$

Assume  $r_2=r_{21}:\dots:r_{2k}$ , form region  $d=MBR(b) \cap r_{21}(c) \cap \dots \cap r_{2k}(c)$  satisfying  $MBR(b)=MBR(d)$ . Therefore,

$$dir(a, d)=r_1 \quad dir(d, c)=r_2 \quad dir(a, c)=t$$

so,  $t \in r_1 \circ r_2$  is satisfied.

**Lemma 1:** Let  $r_1$  be atomic direction,  $r_2$  be a rectangular direction and  $t \in r_1 \circ r_2$ , then the following implication holds

$$(\forall b, c \in REG^*) (dir(b, c)=r_2 \rightarrow \exists a (dir(a, b)=r_1 \quad dir(a, c)=t)).$$

**Proof:** If  $\forall b, c \in REG^*$  satisfying  $dir(b, c)=r_2$ , then assume  $\beta_1(c), \dots, \beta_9(c)$  are the 9 tiles divided by  $MBR(c)$ , it must exist a maximal subset  $\{\delta_i(c), \dots, \delta_m(c)\} \subseteq \{\beta_1(c), \dots, \beta_9(c)\}$ , such that  $i \in \{1..m\}$ ,

$$\delta_i(c) \cap r_l(b) \neq \emptyset \forall \delta_j(c) \in \{\beta_1(c), \dots, \beta_9(c)\} \setminus \{\delta_1(c), \dots, \delta_m(c)\} \quad \delta_j(c) \cap r_l(b) = \emptyset.$$

Let  $t = t_1 : \dots : t_k$  and  $\delta = \delta_1 : \dots : \delta_m$ , if  $t \in r_1 \circ r_2$  then  $t \subseteq \delta$  holds, i.e.,

$\exists a \in REG^*$  such that  $a \in r_l(b) \quad a \cap t_i(c) \neq \emptyset, 1 \leq i \leq k$ .

Otherwise, there exists  $t_p, 1 \leq p \leq k$  such that

$$t_p(c) \in \{\beta_1(c), \dots, \beta_9(c)\} \setminus \{\delta_1(c), \dots, \delta_m(c)\} \quad a \cap t_p(c) \neq \emptyset$$

holds. Since  $a \in r_l(b)$  then  $r_l(b) \cap t_p(c) \neq \emptyset$  which contradicts  $r_l(b) \cap t_p(c) = \emptyset$ . ■

To facilitate the illustration,  $\sigma(s_1, \dots, s_m)$  is the shortcut for the set of all valid basic directions that can be constructed by cross joining the set  $s_1, \dots, s_m$ , such as  $\sigma(\{O\}, \{W: NW, W\}) = \{O: W: NW, O: W\}$ .

**Theorem 2:** Let  $r_1 = r_{1l} : \dots : r_{1k}$  be basic direction and  $r_2$  be rectangular direction then  $r_1 \circ r_2 = (r_{1l} \circ r_2, \dots, r_{1k} \circ r_2)$  holds.

**Proof:** Let  $t \in r_1 \circ r_2$ , there exist  $a, b, c \in REG^*$  such that

$$dir(a, b) = r_1 \quad dir(b, c) = r_2 \quad dir(a, c) = t.$$

Since  $dir(a, b) = r_1$ , there exist  $a_1, \dots, a_k \in REG^*$  such that

$$a = a_1 \dots a_k \text{ and } dir(a_1, b) = r_{11} \dots dir(a_k, b) = r_{1k} \quad dir(b, c) = r_2 \quad dir(a, c) = t$$

Therefore, there exist basic relations  $t_1, \dots, t_k$  and  $t \in \sigma(t_1, \dots, t_k)$  such that  $dir(a_i, c) = t_i, 1 \leq i \leq k$  are satisfied. So

$$dir(a_i, b) = r_{1i} \quad dir(b, c) = r_2 \quad dir(a_i, c) = t_i \text{ i.e., } t_i \in r_{1i} \circ r_2.$$

It means  $t \in \sigma(r_{11} \circ r_2, \dots, r_{1k} \circ r_2)$ .

Conversely,  $\forall t \in \sigma(r_{11} \circ r_2, \dots, r_{1k} \circ r_2)$ , assume  $t = \sigma(t_1, \dots, t_k)$  and  $t_i \in r_{1i} \circ r_2, 1 \leq i \leq k$ . There exist  $a_i, b, c \in REG^*$  such that

$$dir(a_i, b) = t_{1i} \quad dir(b, c) = r_2 \quad dir(a_i, c) = t_i$$

by Lemma 1. Therefore, let  $a = \bigcup_i a_i$ , then

$$dir(a, b) = r_1 \quad dir(b, c) = r_2 \quad dir(a, c) = t$$

is satisfied, i.e.,  $t \in r_1 \circ r_2$ . ■

**Theorem 3:** Let  $r_1$  be an atomic direction and  $r_2$  be a rectangular direction, then

$$r_1 \circ r_2 = \{ \bigcup_i ORG(t) : t \in CD(REC(r_1) \circ REC(r_2)) \}$$

**Proof:**  $\forall u \in r_1 \circ r_2$ , there must exist  $a, b, c \in REG^*$ , such that

$$dir(a, b) = r_1 \quad dir(b, c) = r_2 \quad dir(a, c) = u$$

Since  $r_1$  and  $r_2$  are rectangular,

$$dir(MBR(a), MBR(b)) = r_1 \quad dir(MBR(b), MBR(c)) = r_2$$

From the point of rectangle algebra

$$MBR(a) REC(r_1) MBR(b) \quad MBR(b) REC(r_2) MBR(c)$$

So  $dir(MBR(a), MBR(c)) \in REC(r_1) \circ REC(r_2)$ , i.e.,  $SRD(u) \in CD(REC(r_1) \circ REC(r_2))$ .

Let  $t=SRD(u)$  then  $u \in ORG(t)$ .

Conversely,  $\forall t \in CD(REC(r_1) \circ REC(r_2))$ , there must exist rectangle  $a, b$  and  $c \in REG^*$  such that

$$dir(a, b)=r_1 \quad dir(b, c)=r_2 \quad dir(a, c)=t.$$

Then assume  $\forall u=u_1; \dots; u_k \in ORG(t)$ , form a rectangle  $a_0=a \cap u_1(c) \cap \dots \cap u_k(c)$ , we have

$$dir(a_0, b)=r_1 \quad dir(b, c)=r_2 \quad dir(a_0, c)=u.$$

Therefore  $u \in r_1 \circ r_2$  holds.

To sum up, for arbitrary basic relation  $r_1=r_{11}; \dots; r_{1k}$  and  $r_2$ , following equations must hold:

$$r_1 \circ r_2 = r_1 \circ SRD(r_2)$$

$$r_1 \circ r_2 = \sigma(r_{11} \circ SRD(r_2), \dots, r_{1k} \circ SRD(r_2))$$

$$r_1 \circ r_2 = \sigma(ORG(t_1), \dots, ORG(t_k)) \quad \text{where } t_i \in CD(REC(r_{1i}) \circ REC(SRD(r_2))),$$

$1 \leq i \leq k$ .

Then we get following composing algorithm *CDCom*

---

**Algorithm** CDComp

**Input:** basic cardinal direction  $r, t \in D^*$

**Output:** the result of  $r \circ t$ , the set  $R$

1. Let  $q=SRD(t)$ , compute the corresponding rectangle relation  $REC(q)$
2. Assumer= $r_1; \dots; r_k$ , compute the rectangle relation of  $r_i$ ,  $REC(r_i)$ , where  $i=1, \dots, k$   
For  $i=1, \dots, k$

Compute  $P_i = REC(r_i) \circ REC(q)$  and  $CD(P_i)$

3. For  $i=1, \dots, k$

$$T_i = \emptyset$$

For each  $d \in CD(P_i)$

$$\text{Compute } T_i = T_i \cup ORG(d)$$

4. Return  $R = \sigma(T_1, \dots, T_k)$ . ■
- 

**Example 4.** Compute the composition of  $r=W:O, t=W:SW:S$

1. Let  $q=SRD(t)=W:O:SW:S$ , then  $REC(q)=\{o, fi\} \times \{o, fi\}$  according to Table 3.
2. The atomic directions  $r$  including are  $W$  and  $O$ , then we have:

$$REC(W)=\{p, m\} \times \{e, d, s, f\},$$

$$P_1=REC(W) \circ REC(q)=\{p\} \times \{p, m, o, fi, e, d, s, f\}$$

$$CD(P_1)=\{SW, W:SW, W\}$$

$$REC(O)=\{e, d, s, f\} \times \{e, d, s, f\}$$

$$P_2=REC(O) \circ REC(q)=\{p, m, o, fi, e, d, s, f\} \times \{p, m, o, fi, e, d, s, f\},$$

$$CD(P_2)=\{SW, W:SW, W, SW:S, SW:S:O:W, W:O, S, S:O, O\}$$

3. From Theorem 3, we have:

$$T_1=\{SW, W:SW, W\},$$

$$T_2=\{SW, W:SW, W, SW:S, SW:S:O:W, S:O:W, SW:O:W, SW:S:W, SW:S:O:W:O, S, S:O, O, W:S, O:SW\}$$

4. The final result  $R = \sigma(T_1, T_2) = \{W:O:SW:S, W:O:SW, W:O:S, W:SW:S, O:SW:S, W:O, SW:S, W:SW, W, SW, SW:O, W:S\}$  as Fig.3 shows orderly.

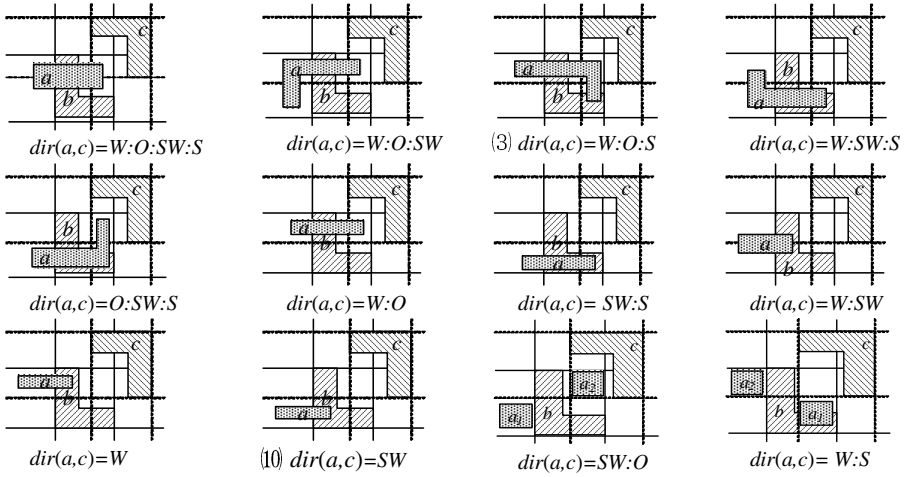


Fig. 3. Pictorial examples of  $W:O \circ W:SW:S$

### 6. Conclusions

Basing on the concepts of smallest rectangular directions and the original directions in SK-model, this paper translates the composition of basic cardinal directions into the composition of interval relations and gives the proof. Comparing with existing method, the given algorithm has quite good dimensional extendibility. The similar results can be easily attained in tridimensional space just like the extension from interval algebra to block algebra [18]. While two issues should be mentioned here are that algorithm *CDComp* can only process the consistency-based composition and work correctly over compound regions *REG\**.

Algorithm *CDComp* gives consistency-based composition (weak composition), which can be easily derived from the proofs of Theorem 1~4. For arbitrary direction  $r$  and  $s$ :  $r \circ_w s = \{t_1, \dots, t_m\}$  where  $\forall a \forall b \forall c ((dir(a, b)=r \wedge dir(b, c)=s) \rightarrow (dir(a, c)=t_1 \vee \dots \vee dir(a, c)=t_m))$  holds.  $N \in N \circ_w N$ , as the left figure of Fig. 4 shows there exist regions satisfy the above formula. While for the existence-based composition (strong composition),  $\forall dir(a, c) \in r \circ_s s \leftrightarrow \exists b: dir(a, b)=r \wedge dir(b, c)=s$ .  $N \notin N \circ_s N$ , for  $a$  meets  $c$  from north, no region  $b$  satisfy  $dir(a, b)=N \wedge dir(b, c)=N \wedge dir(a, c)=N$  as the right figure shows.

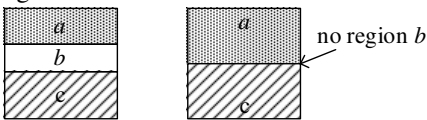


Fig. 4. explanation of weak composition.

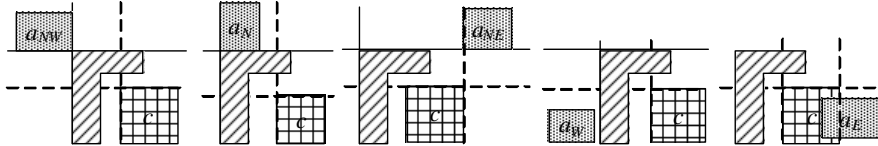
Algorithm *CDComp* can only give correct results over *REG\**. The main reason lies in the proof of Theorem 2. When forming new region  $a = \dots a_i$ , there is no condition to

limit  $a$  as a simple region, so it can only ensure  $a$  be a compound region. Example 6 gives the instance. According to Theorem 3

$$\begin{aligned} NW \in NW \circ NW:N:W \quad NW \in N \circ NW:N:W, \quad NE \in NE \circ NW:N:W \\ W \in W \circ NW:N:W, \quad O:E \in E \circ NW:N:W \end{aligned}$$

as shown in Fig. 5, therefore we have  $NW:NE:W:O:E \in NW:N:NE:W:E \circ NW:N:W:W:E$   $dir(b, c)=NW:N:W$   $dir(a, c)=NW:NE:W:O:E$  holds.

**Example 6.** let  $r=NW:N:NE:W:E$ ,  $t=NW:N:W$  then  $s=NW:NE:W:O:E \in r \circ t$  can only be satisfied in  $REG^*$



**Fig. 5.** illustration of evidences of example 6

Most potential applications of QSR involve different aspects of space. So reasoning with multi-aspect spatial information has become the focus of QSR. Some work has been done in this issue [11]-[15] but it mainly concentrates on the two aspects of space such as topology and direction, topology and size, distance and direction. It lacks the integrative reasoning over three or more aspects of spatial relations. Due to the good extendibility of interval algebra which can represent both directional and topological information in one calculus; it is worth introducing other type information into it to realize the integrative reasoning over three or more aspects of spatial relations.

## Acknowledgement

This paper is supported by National Natural Science Foundation of China under Grant Nos. 60773099, 60873149, 60973088, the National High-Tech Research and Development Plan of China under Grant 2006AA10Z245, 2006AA10A309, Research Fund for the Doctoral Program of Higher Education of China under Grant No 20070783057, Special funds of Central Colleges Basic Scientific Research Operating Expenses, Jilin University under Grant No. 421032041421.

Contact author: Haiyang Jia, [jjahy@jlu.edu.cn](mailto:jjahy@jlu.edu.cn), Jilin University, 2699 Ave. Qianjin, Changchun, 130012, P.R.China.

## References

- [1] Cohn, A.G., Hazarika, S.M.: Qualitative spatial representation and reasoning: An overview. *Fundamenta Informaticae* 46(1), 2–32 (2001)
- [2] Egenhofer, M.J.: Spherical Topological Relations. *Journal on Data Semantics* 3(1), 25–49 (2005)
- [3] Renz, J.: Qualitative Spatial Reasoning with Topological Information. LNCS (LNAD), vol. 2293. Springer, Heidelberg (2002)

- [4] Andrew, U.F.: Qualitative spatial reasoning about cardinal directions. In: Proc. of 7th Austrian Conference on Artificial Intelligence, pp. 157–167. Morgan Kaufmann, Baltimore (1991)
- [5] Christian, F.: Using orientation information for qualitative spatial reasoning. In: Frank, A.U., Formentini, U., Campari, I. (eds.) GIS 1992. LNCS, vol. 639, pp. 162–178. Springer, Heidelberg (1992)
- [6] Goyal, R.K., Egenhofer, M.J.: Consistent Queries over Cardinal Directions across Different Levels of Detail. In: Proc. of 11th Int. Workshop on Database and Expert Systems Applications, Greenwich, London, UK, pp. 876–880. IEEE Press, Los Alamitos (2000)
- [7] Goyal, R.K., Egenhofer, M.J.: Similarity of Cardinal Directions. In: Jensen, C.S., Schneider, M., Seeger, B., Tsostras, V.J. (eds.) SSTD 2001. LNCS, vol. 2121, pp. 36–55. Springer, Heidelberg (2001)
- [8] Skiadopoulos, S., Koubarakis, M.: Composing cardinal direction relations. *Artificial Intelligence* 152(1), 143–171 (2004)
- [9] Skiadopoulos, S., Koubarakis, M.: On the consistency of cardinal direction constraints. *Artificial Intelligence* 163(1), 91–135 (2005)
- [10] Cicerone, S., Felice, P.D.: Cardinal directions between spatial objects: the pairwise-consistency problem. *Information Sciences* 164(1), 165–188 (2004)
- [11] Clementini, E., Felice, P., Hernandez, D.: Qualitative representation of positional information. *Artificial Intelligence* 95(2), 317–356 (1997)
- [12] Liu, J.: A method of spatial reasoning based on qualitative trigonometry. *Artificial Intelligence* (1-2), 137–168 (1998)
- [13] Sistla, A.P., Bu, C.: Reasoning about qualitative spatial relationships. *Journal of Automated Reasoning* 25(4), 291–328 (2000)
- [14] Gerevini, A., Renz, J.: Combining topological and size information for spatial reasoning. *Artificial Intelligence* 137(1), 1–42 (2002)
- [15] Li, S.: Combining Topological and Directional Information for Spatial Reasoning. In: Proc. of Int. Joint Conf. on Artificial Intelligence, Hyderabad, India, pp. 435–440 (2007)
- [16] Allen, J.: Maintaining Knowledge about Temporal Intervals. *Communications of the ACM* 26(11), 832–843 (1983)
- [17] Balbiani, P., Condotta, J.-F., Fariñas del Cerro, L.: A new Tractable Subclass of the Rectangle Algebra. In: Proc. of Int. Joint Conf. on Artificial Intelligence, Stockholm, Sweden, pp. 442–447 (1999)
- [18] Balbiani, P., Condotta, J.-F., Fariñas del Cerro, L.: Tractability results in the block algebra. *Journal of Logic and Computation* 12(5), 885–909 (2002)

# Retrieval Result Presentation and Evaluation

Shengli Wu<sup>1</sup>, Yaxin Bi<sup>1</sup>, and Xiaoqin Zeng<sup>2</sup>

<sup>1</sup> School of Computing and Mathematics  
University of Ulster, Northern Ireland, UK

{s.wu1,y.bi}@ulster.ac.uk

<sup>2</sup> School of Computer Science  
Hohai University, Nanjing, China  
xzeng@hhu.edu.cn

**Abstract.** In information retrieval systems and digital libraries, result presentation is a very important aspect. In this paper, we demonstrate that only a ranked list of documents, thought commonly used by many retrieval systems and digital libraries, is not the best way of presenting retrieval results. We believe, in many situations, an estimated relevance probability score or an estimated relevance score should be provided for every retrieved document by the information retrieval system/digital library. With such information, the usability of the retrieval result can be improved, and the Euclidean distance can be used as a very good system-oriented measure for the effectiveness of retrieval results. The relationship between the Euclidean distance and some ranking-based measures are also investigated.

## 1 Introduction

In information retrieval systems and digital libraries, retrieval result presentation and evaluation are two important but related issues. The commonest format for retrieval result presentation is a single list of documents, in which all the documents are ranked in a way that their estimated relevancies decrease from the highest to the lowest. Then relevance judgment such as binary relevance judgment or graded relevance judgment can be used to evaluate its effectiveness. In binary relevance judgment, documents are classified into two categories: relevant or irrelevant. In graded relevance judgment, documents are divided into  $n + 1$  categories: grades  $n, n - 1, \dots, 0$  ( $n \geq 2$ ). The documents in grade  $n$  are the most relevant, which are followed by the documents in grade  $n - 1, n - 2, \dots, 1$ , and the documents in grade 0 are irrelevant. In the evaluation of retrieval results, precision (the ratio of relevant documents retrieved to all retrieved documents) and recall (the ratio of relevant documents retrieved to all relevant documents in the whole collection) are the two most important measures. Recently, some of their derivations such as average precision over all relevant documents, recall-level precision, and average precision at given document cutoff values, have been widely used for retrieval evaluation.

In this paper, we demonstrate that only a single list of documents as retrieval result is not enough in some situations. Let us consider two examples. The first

is for the application of patent and legal case retrieval, it may require that all relevance documents need to be retrieved. The second one is that a researcher wants to find all articles which are relevant to his/her research topic. In both situations, if the retrieval system only provides a large number of ranked documents for a given information need, then the user may feel very frustrated. First, unless the user checks all the documents in the result, one has no idea how many relevant documents are in the whole list. Second, if the user plans to find a given number (say, 10) relevant documents, one has no idea how many documents needs to be checked in the list. Third, one has no idea if all the relevant documents have been included in the result or not.

As a matter of fact, it is a difficult task for existing information retrieval systems and digital libraries to provide a precise answer to the above questions. However, some reasonable estimations, which can be managed to provide by many information retrieval systems and digital libraries, are also very useful. If a large amount of documents need to be checked, then it is better to know (at least roughly) the number of them before starting to do so. Then an appropriate period of time can be arranged for that.

In order to solve the above problems, more information is needed for retrieval result presentation. In this paper we discuss two related solutions. One solution is to provide an estimated probability score for every document retrieved. This solution is applicable when binary relevance judgment is used. The second solution is to provide an estimated relevance score for every document retrieved. This solution is applicable when graded relevance judgment is used. With these scores, the users can have a clearer view of the result; it is also possible to give a reasonable answer to the above questions.

The rest of this paper is organized as follows: in Section 2 we review some related work. Relevance probability scores and relevance scores are introduced in Section 3 and Section 4, respectively. The Euclidean distance is discussed in Section 4. The relationship between the Euclidean distance and other ranking based measures is discussed in Section 5. Section 6 is the conclusion.

## 2 Related Work

In information retrieval, result evaluation is an important and complicated task. First of all, “relevance” is a central but rather equivocal concept [1, 15, 16] since deciding if a document is relevant to an information need is not totally objective. Harter [6] divided those factors that might affect relevance judgment into six groups: characteristics of judges, queries, documents, information systems, judgment conditions, and choice of scale. Second, relevance judgment is a task which demands huge human effort. In some situations such as evaluating searching services on the World Wide Web, complete relevance judgment is not possible. It is also not affordable when using large document collections for the evaluation of information retrieval systems. For example, in the Text REtrieval Conferences (TREC)<sup>1</sup> held by the National Institute of Standards and Technology of the USA,

---

<sup>1</sup> Its home web page is located at <http://trec.nist.gov/>



only partial relevance judgment was conducted due to the large number of documents (from 0.5 to over 10 million) in the whole collection. A pooling method [8] has been used in TREC. Its effect on evaluation has been investigated in [2, 14, 21, 22] among others.

Binary relevance judgment has been used for long time [19]. Precision and recall are defined in such a circumstance. So do many other measures [18] which are derived from precision and recall (such as average precision over all relevant documents, recall-level precision, average precision at given document cutoff values). Järvelin and Kekäläinen [7] proposed several measures including cumulated gain, discounted cumulated gain, and normalized discounted cumulated gain for graded relevance judgment. Wu and McClean [21] extended the definitions of average precision over all relevant documents and recall-level precision for graded relevance judgment. In TREC [18], most tracks used both binary relevance judgment, but some of them used three category relevance judgment (irrelevant, modestly relevant, and highly relevant).

To present retrieval result, a single list of ranked documents is the commonest option. An alternative is to present retrieval results by using hierarchical structures. Some previous research [5, 17] suggests it can be as good as the single-list option.

When ranking documents for a given information need, most information retrieval systems/digital libraries assign a retrieval status value to every document in the document collection, then rank all the documents according to the status values they obtain. It should be noted that such retrieval status values usually do not necessarily have to be the (estimated) relevance probabilities of the documents. As a consequence, little effort has been made on approximating the relationship between retrieval status values and relevance probabilities [12].

In some advanced IR applications such as filtering, resource selection and data fusion, people find that relevance probabilities of documents are highly desirable. Therefore, different kinds of score normalization methods, which map internal retrieval status values to probabilities of relevance, have been proposed and their effectiveness has been investigated in [3, 9, 10, 11, 12].

To the best of our knowledge, using relevance probabilities to improve the usability of information retrieval systems/digital libraries has not been paid much attention before. This is one of the major issues which will be discussed in this paper. Furthermore, we will discuss relevance scores and the exact relevance judgment for retrieval results. Some good properties of them will be discussed. Although the same idea may also be useful for the hierarchical result presentation, we only consider the situation of single-list presentation in this paper.

### 3 Relevance Probability Scores

As mentioned in Section 2, information retrieval systems use retrieval status values to rank documents, but such retrieval status values usually do not necessarily have to be relevance probabilities or relevance degrees of the documents to the information need. Therefore more effort is needed if we want to estimate the probabilities of relevance or degrees of relevance of those retrieved

documents. However, how to provide such scores by information retrieval systems/digital libraries is beyond the score of this paper, though it is an important issue (see [3,10,11,12] and others for related discussion). Instead, we discuss why we need them and what we can do by using them.

First of all, such scores are very desirable for various IR applications such as distributed information retrieval systems/digital libraries, data fusion and others. In these applications, for every document, its global score is calculated based on its scores achieved in all component retrieval systems. If considering a standalone information retrieval system/digital library, such scores for retrieved documents are very useful as well. In the rest of this section, let us assume that relevance probability scores are available for all retrieved documents and binary relevance judgment is applied, we discuss how to use them to improve retrieval result's usability.

For a query (information need)  $Q$ , an information retrieval system  $S$  retrieves a group of documents  $D=\{d_1,d_2,\dots,d_n\}$ , each of which ( $d_i$ ) is assigned a corresponding relevance probability score  $s_i$  for ( $1 \leq i \leq n$ ). Further, we assume that these documents have been ranked according to their scores. Therefore, we have  $s_1 \geq s_2 \geq \dots \geq s_n$ . Now we are in a position to answer those questions raised in Section 1. First let us discuss how to answer the first question: how many relevant documents are in the list?

Obviously, as a point estimate we can expect  $\sum_{i=1}^n s_i$  relevant documents in resultant list  $D$ . We can also estimate the probabilities that  $D$  includes 0,1,..., $n$  relevant documents are:

$$\begin{aligned}
 p(0) &= \prod_{i=1}^n (1 - s_i) \\
 p(1) &= \sum_{i=1}^n [ \prod_{j=1 \wedge j \neq i}^n (1 - s_j) ] s_i \\
 &\dots \\
 p(k) &= \sum_{i=1}^{\frac{n!}{k!(n-k)!}} [ ( \prod_{j=1}^k s_{ij} ) ( \prod_{j=1}^{n-k} (1 - s'_{ij}) ) ] \\
 &\dots \\
 p(n) &= \prod_{i=1}^n s_i
 \end{aligned}$$

Here  $p(k)$  ( $1 \leq k \leq n$ ) is the probability that resultant list  $D$  includes  $k$  relevant documents. We explain a little more about  $p(k)$ . For  $k$  relevant documents in a list of  $n$  ( $n > k$ ) documents, there are  $\frac{n!}{k!(n-k)!}$  different combinations. In each of these combinations  $c_i$ , the scores of  $k$  relevant documents are denoted as  $s_{i1}, s_{i2}, \dots, s_{ik}$ , and the scores of  $n - k$  irrelevant documents are denoted as  $s'_{i1}, s'_{i2}, \dots, s'_{i(n-k)}$ . With these probability values, we can answer a series of

questions. For example, the probability that resultant list  $D$  includes at least  $k$  relevant documents is  $\sum_{i=k}^n p(i)$ .

The second question is: how many documents do we need to check if we want to obtain  $k$  ( $k > 0$ ) relevant documents? Obviously it is better to check those documents in the same order as they are ranked. The following estimation is based on such an assumption. In order to obtain  $k$  relevant documents, the expected number of documents we need to check is  $m$ , where  $\sum_{i=1}^m s_i > k$  and  $\sum_{i=1}^{m-1} s_i < k$ .

The third question is: how many relevant documents are not in the resultant list? In order to answer this question, we need to estimate the relevance probabilities of those documents which are not in the resultant list. This can be estimated by using the relevance probability scores of the resultant list. For example, logistic functions [3] and cubic functions [20] have been found useful for such a purpose. We shall illustrate how to do this in Example 2 later in this section.

All those measures, such as average precision over all relevant documents, recall-level precision, and average precision at given document cutoff values, can be used for effectiveness evaluation as before, because a ranked list of documents is still the major part of the retrieval result. However, for those scores associated with documents, we can use some measures to evaluate their accuracy. For example, we may define  $Ratio = \frac{\sum_{i=1}^n s_i}{N_r}$ , where  $N_r$  is the judged real number of relevant documents in result  $D = \{d_1, d_2, \dots, d_n\}$  and  $s_i$  is the score possessed by  $d_i$ . If  $Ratio$  is bigger than 1, it means that those probability scores have been overestimated; if  $Ratio$  is smaller than 1, it means that those probability scores have been underestimated. No doubt, 1 is the ideal value for  $Ratio$ . In addition,  $Ratio$  does not necessarily be defined for the whole document list, it can also be defined for a subset of documents, for example, a group of top-ranked documents, in the resultant list.

**Example 1.** A resultant list includes 12 documents. Their scores are  $\{0.60, 0.56, 0.52, 0.48, 0.44, 0.40, 0.36, 0.32, 0.28, 0.24, 0.20, 0.16\}$ .

A point estimate of the number of relevant documents in this list is

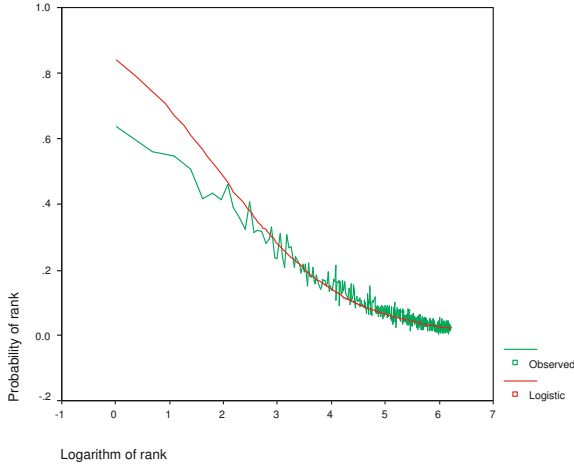
$$\sum_{i=1}^{12} s_i = 4.56 \approx 5$$

If we want to find 2 relevant documents, a point estimate of the number of documents we need to check is 4 since

$$\sum_{i=1}^4 s_i = 2.16 > 2$$

and

$$\sum_{i=1}^3 s_i = 1.68 < 2$$



**Fig. 1.** One submitted result (input.fub04De) to TREC 2004 (robust track)

**Example 2.** Let us consider one submitted run (input.fub04DE), which was submitted to the TREC 2004 robust track. 1000 documents were retrieved for each query. In this result, the relationship between the probability of relevance and the logarithm of rank over a total of 249 queries for the top-500 documents is shown in Fig. 1 as the observed curve (the one with many twists). We can estimate the observed curve by some other curves such as linear, logistic [3], etc. The logistic model uses the following function

$$p = \frac{1}{1 + a * e^{\ln(rank) * \ln(b)}} \quad (1)$$

to estimate curves. Using the top-500 documents to generate the estimated curve, then we use the same curve to estimate the number of relevant documents in the bottom-500 documents, and compare the estimated numbers with real numbers to see how accurate the estimation is. The estimated curve of input.fub04De using the logistic model is shown in Fig. 1 as the logistic curve (the smooth curve). we obtain the values for coefficients  $a$  and  $b$  in Equation 1. They are 0.1869 for  $a$  and 2.3827 for  $b$ . Therefore, we can estimate the relevance probability of a document at ranks 501-1000 using Equation 1 with the above parameters. As a step further, we can estimate the number of relevant documents for those documents ranking from 501 to 1000. The estimated number is 8.63 while the real number is 7.56 (a total of 1882 for 249 queries) for input.fub04De. Besides this run, we randomly chose 8 other submitted runs and compared their estimated numbers with real numbers of relevant documents at ranks 500-1000, which are shown in Table 1. It seems that the estimation is quite accurate and the logistic function is a good function for this purpose.

**Table 1.** Estimated number vs. real numbers of relevant documents at ranks 501-1000 for 8 randomly chosen submitted runs in TREC 2004 (error= | estimated-real | /real)

Submitted run	Estimated	Real	Error
apl04rsTDNfw	8.31	7.74	7.36%
apl04rsTDNw5	6.48	6.29	3.02%
humR04d4e5	7.31	6.09	20.03%
icl04pos2d	7.44	7.51	0.93%
JuruDes	7.19	6.53	10.11%
uogRobSWR10	9.24	7.24	27.62%
vrumtitle	8.73	7.25	20.41%
wdoqla1	8.03	7.33	9.55%

## 4 Relevance Scores

Binary relevance judgment has been widely used in various information retrieval evaluation activities such as TREC [18], CLEF [4], NTCIR [13], etc. This is mainly due to the fact that less human effort is required when using binary relevance judgment. For any document, only two options are available for judges to choose from: relevant or irrelevant. A threshold of 50% can be set. If the relevance degree of a document is 50% or over, then it is regarded as a relevant document; otherwise, it is regarded as an irrelevant document. However, in some situations, a more accurate relevance judgment is suitable. For example, if  $n$  referees are involved in the judgment of a document and all of them use binary relevance judgment, then there are  $n + 1$  possibilities: 0 or 1 or ... or  $n$  referees may judge the document to be relevant. In such a situation, binary relevance judgment can be used for making the final decision: if more referees are in favour of the decision of being relevant, then the final decision is “relevant”; otherwise, the final decision is “irrelevant”. It is better to use an odd number of referees for the judgment. Otherwise, it may happen that equal number of referees are in favour of and against the decision of being relevant, then no appropriate decision can be made. However, this solution is very obscure. For a “relevant” or “irrelevant” document, we do not know how many or what percentage of referees favour such a decision. A more accurate solution is to use  $n + 1$  graded relevance judgment. If  $m$  ( $0 \leq m \leq n$ ) referees are in favour of the “relevant” decision, then the document is in grade  $m$ . ( $n + 1$ ) graded relevance judgment has a resolution of  $n + 1$ , while binary relevance judgment only has a resolution of 2.

Theoretically, if we do not consider the cost of human judgment, the exact relevance judgment is the best since it has a resolution of an indefinite large number. Graded relevance judgment and binary relevance judgment can be regarded as approximations of the exact relevance judgment. The more grades we use in graded relevance judgment, the more approximate we are to the exact relevance judgment. Binary relevance judgment is the roughest approximation of the exact relevance judgment.

If every document retrieved bears a relevance score, then the exact relevance judgment, graded relevance judgment, or binary relevance judgment can be applied. These scores may look a little different from relevance probability scores, which we used in Section 3. It makes sense to use relevance probability scores with binary relevance judgment only, but relevance scores can be used by different kinds of relevance judgments including binary relevance judgment. One hypothesis is: for the same scores in the range of  $[0,1]$ , we can explain them in either way: probability of relevance or degree of relevance.

If scores of relevance degree are provided, we can use the Euclidean distance as a measure to evaluate the effectiveness of any result. For a result  $D = \{d_1, d_2, \dots, d_n\}$ ,  $S = \{s_1, s_2, \dots, s_n\}$  and  $H = \{h_1, h_2, \dots, h_n\}$  are system-provided relevance scores and human-judged relevance scores for  $D$ , respectively. The effectiveness of  $S$  can be defined as the Euclidean distance between  $S$  and  $E$

$$dis(S, H) = \sqrt{\sum_{i=1}^n (s_i - h_i)^2} \quad (2)$$

For example, suppose there are three documents in the result and their system-provided scores and human-judged scores are  $S = \{1.0, 0.7, 0.2\}$  and  $H = \{0.0, 0.4, 0.5\}$ , respectively, then

$$dis(S, H) = \sqrt{(1.0 - 0.0)^2 + (0.7 - 0.4)^2 + (0.2 - 0.5)^2} = 1.09$$

In Equation 2, all the documents in the resultant list contribute equally to the final value of  $dis(S, H)$ . Therefore, the Euclidean distance is a very good system-oriented measure for effectiveness evaluation. Note that the Euclidean distance can be used with graded relevance judgment or binary relevance judgment.  $H = \{h_1, h_2, \dots, h_n\}$  is decided in the following way: if there are  $n + 1$  grades, then for any document in grade  $i$ , the document has a relevance score of  $i/(n + 1)$ .

Next let us discuss some possible variations of the Euclidean distance. Sometimes the users are just interested in a few top-ranked documents in the whole resultant list to see if they are useful or not. This is especially the case when the retrieved result includes a large number of documents, then it is difficult to read all the documents in the result. Therefore, the average precision at given document cutoff values (say  $m=5, 10, \text{ or } 20$ ) is a common user-oriented measure with binary relevance judgment. Accordingly, we may also define the Euclidean distance of the  $m$  top-ranked documents as:

$$dis_{stop-m}(S, H) = \sqrt{\sum_{i=1}^m (s_i - h_i)^2} \quad (3)$$

Another variation of the Euclidean distance can also be defined. Suppose for the information need in question,  $Best_m = \{b_1, b_2, \dots, b_n\}$  are the judged scores of  $m$  most relevant documents in the whole collection and  $H_m = \{h_1, h_2, \dots, h_m\}$  are judged scores of the  $m$  top-ranked documents in the resultant list. Then we may define the distance between  $H_m$  and  $Best_m$

**Table 2.** Query results of two hypothetical retrieval systems ('r' denotes a relevant document and 'i' denotes an irrelevant document)

Query	$R_1$	$R_2$	$\text{dis}(R_1, H)$	$\text{dis}(R_2, H)$	$\text{AP}(R_1)$	$\text{AP}(R_2)$
1	(.9, r; .8, r)	(.3, r; .2, r)	.2236	1.063	1	1
2	(.65, r; .35, i)	(.65, i; .35, r)	.4950	.9192	1	.5
3	(.9, r; .8, i)	(.6, i; .5, r)	.8062	.7810	1	.5

$$\text{dis}_{\text{best}_m}(H_m, \text{Best}_m) = \sqrt{\sum_{i=1}^m (h_i - b_i)^2} \quad (4)$$

This measure is analogous to recall in a sense that the top  $m$  documents in result  $D$  are compared with the  $m$  most relevant documents in the collection.

## 5 Relationship between the Euclidean Distance and Other Measures

In information retrieval, most commonly used system-oriented measures such as average precision over all relevant documents (AP) and recall-level precision (RP) are ranking based measures. This means, when using these measures to evaluate results, only the ranking positions of relevant and irrelevant documents matters (in case of binary relevance judgment), while the relevance scores of the documents are ignored. Compared with those ranking based measures, the Euclidean distance is more accurate. Let us illustrate this by an example. Suppose we have two information retrieval systems  $I_1$  and  $I_2$  and three queries  $q_1$ ,  $q_2$ , and  $q_3$ . For each query, both  $I_1$  and  $I_2$  retrieve two documents with relevance scores. We also assume that all the relevant documents are included in every result for every query. Three hypothetical results are shown in Table 2.

For Query 1, both results include two relevant documents. Therefore, for both results, their AP values are 1. However, their Euclidean distances are very different (.2236 versus 1.063). This is because in the first result, both relevant documents are assigned high relevance scores (0.9 and 0.8); while in the second result, both relevant documents are assigned very low relevance scores (0.3 and 0.2). For Query 2, two measure values are consistent. For Query 3, result  $R_1$  is twice as good as  $R_2$  on AP; however,  $R_1$  is slightly worse than  $R_2$  on the Euclidean distance. From these examples, we can see that the Euclidean distance is a more accurate measure than AP. On the other hand, these two measures may cause significant difference, especially when very few queries are performed. It is interesting to find how these two kinds of measures correlate when a large number of queries are performed. We carried out an experiment to investigate this.

We use statistical methods to empirically investigate this issue by experimenting with a group of submitted runs to the TREC 2001 Web track, in which three grades were used for relevance judgment. All selected runs (32 in total) include 1000 documents for every query.

**Table 3.** Pearsons correlation coefficients among different measures

Measure	AP	RP	NDCG2	P10
Euclidean Distance	-.952	-.950	-.978	-.882
MAP		.963	.973	.875
RP			.966	.904
NDCG2				.902

**Table 4.** Linear regression of different measures (dependent variable is the Euclidean distance)

Measure	Constant	Linear coefficient	$R^2$	Significance level
AP	4.644	-1.243	0.907	< .0005
RP	4.717	-1.284	0.902	< .0005
NDCG2	4.795	-0.855	0.957	< .0005
P10	4.603	-0.854	0.778	< .0005

In TREC 2001, grade 0 means irrelevant, grade 1 means modestly relevant, while grade 2 means highly relevant documents. We use 0, 0.5, and 1 to score irrelevant, modestly relevant, and highly relevant documents in qrels (relevance judgment document), respectively. In this experiment we used the cubic model [20] (the logistic model [3] is another option). For all selected component results put together (also called runs) over 50 queries, a cubic model  $S(r) = a_0 + a_1 \ln(r) + a_2 \ln(r)^2 + a_3 \ln(r)^3$  was used to estimate relevance scores from ranking information. Here  $r$  is the rank position and  $S(r)$  is the relevance score calculated. The coefficients we obtained from regression analysis are:  $a_0 = 0.4029$ ,  $a_1 = 0.0552$ ,  $a_2 = 0.0026$ , and  $a_3 = 0.0001$  (SPSS was used).

Apart from the Euclidean distance, we used NDCG (normalized discount cumulated gain), AP, RP, and P10 (precision at 10 document level). NDCG was introduced by Järvelin and Kekäläinen in [7] for graded relevance judgment. When using NDCG, parameter  $b$  needs to be set. In our experiment we set  $b=2$  as in [7], and NDCG with such a setting is referred to as NDCG2 later in this paper. Initially, AP, RP, and P10 were used with binary relevance judgment. In this experiment, we used an extension of them, which was defined in [21] for graded relevance judgment. We evaluated the effectiveness of all selected results over 50 queries using all these measures. Then Pearsons correlation coefficients were calculated among different measures, as shown in Table 3.

All the coefficients between the Euclidean distance and other measures are negative since the shorter Euclidean distance is more effective, which is opposite to the four other measures in this respect. In all the cases, the correlations are significant at the 0.01 level (2-tailed). We also carried out a linear regression analysis for those values of different measures. Table 4 shows the coefficients and significance of the analysis.

The Euclidean distance can be well linearly expressed using any of the four other measures. Among them, NDCG2 is the best ( $R^2 = 0.957$ , which means that NDCG2 can explain 95.7% of the variation in the Euclidean distance and vice



versa) and P10 ( $R^2=0.778$ ) is the least able to express the Euclidean distance. Therefore, the experiment demonstrates that, when a relatively large group of queries (say, 50, as in this experiment) are used, then we can use those ranking based measures (AP, RP, NGCG2, P10) to get almost exact evaluation result as using the Euclidean distance. One may wonder why this can be very different from the situation in which an individual query is considered (recall the 3 examples at the beginning of this section). The explanation for this phenomenon is: for a single query, using a ranking based measure usually cannot accurately estimate the effectiveness as the Euclidean distance does. Sometimes it overestimates the effectiveness; some other times it underestimates the effectiveness. However, when a relatively large group of queries are considered, then those underestimates and overestimates cancel each other out and the final average becomes very accurate. Therefore, we can expect that the conclusions drawn from the Euclidean distance are held to a great extent for ranking based measures if a large group of queries are used for the evaluation.

## 6 Conclusion

In this paper we have discussed two related aspects in information retrieval systems and digital libraries: retrieval result presentation and evaluation. They have a certain relationship: retrieval result evaluation can only be carried out when the format of retrieval result presentation has been decided.

We have demonstrated that a single list of documents as retrieval result is not enough in some situations and relevance probability scores or relevance scores should be accompanied with all the documents retrieved. With these scores, the usability of the retrieval results can be improved considerably.

If relevance scores are provided for all the documents retrieved, then the Euclidean distance is a more precise system-oriented measure than those ranking based measures and its use for retrieval result evaluation should be encouraged.

## References

1. Barry, C.L.: User-defined relevance criteria: an exploratory study. *Journal of the American Society for Information Science* 45(3), 149–159 (1994)
2. Buckley, C., Voorhees, E.M.: Retrieval evaluation with incomplete information. In: *Proceedings of ACM SIGIR Conference, Sheffield, United Kingdom, July 2004*, pp. 25–32 (2004)
3. Calvé, A.L., Savoy, J.: Database merging strategy based on logistic regression. *Information Processing & Management* 36(3), 341–359 (2000)
4. CLEF, <http://www.clef-campaign.org/>
5. Crestnai, F., Wu, S.: Testing the cluster hypothesis in distributed information retrieval. *Information Processing & Management* 42(5), 1137–1150 (2006)
6. Harter, S.P.: Variations in relevance assessments and the measure of retrieval effectiveness. *Journal of the American Society for Information Science* 47(1), 37–49 (1996)

7. Järvelin, K., Kekäläinen, J.: Cumulated gain-based evaluation of IR techniques. *ACM Transactions on Information Systems* 20(4), 442–446 (2002)
8. Sparck Jones, K., van Rijsbergen, C.: Report on the need for and provision of an “ideal” information retrieval test collection. Technical report, British library research and development report 5266, Computer laboratory, University of Cambridge, Cambridge, UK (1975)
9. Lee, J.H.: Analysis of multiple evidence combination. In: *Proceedings of the 20th Annual International ACM SIGIR Conference*, Philadelphia, Pennsylvania, USA, July 1997, pp. 267–275 (1997)
10. Manmatha, R., Rath, T., Feng, F.: Modelling score distributions for combining the outputs of search engines. In: *Proceedings of the 24th Annual International ACM SIGIR Conference*, New Orleans, USA, September 2001, pp. 267–275 (2001)
11. Montague, M., Aslam, J.A.: Relevance score normalization for metasearch. In: *Proceedings of ACM CIKM Conference*, Berkeley, USA, November 2001, pp. 427–433 (2001)
12. Nottelmann, H., Fuhr, N.: From retrieval status values to probabilities of relevance for advanced ir applications. *Information Retrieval* 6(3-4), 363–388 (2003)
13. NTCIR, <http://research.nii.ac.jp/ntcir/>
14. Sanderson, M., Zobel, J.: Information retrieval system evaluation: Effort, sensitivity, and reliability. In: *Proceedings of ACM SIGIR Conference*, Salvador, Brazil, August 2005, pp. 162–169 (2005)
15. Saracevic, T.: Relevance: A review of and a framework for thinking on the notion in information science. *Journal of the American Society for Information Science* 26(6), 321–343 (1975)
16. Schamber, L., Eisenberg, M.B., Nilan, M.S.: A re-examination of relevance: toward a dynamic, situational definition. *Information Processing & Management* 26(6), 755–776 (1990)
17. Tombros, A., Villa, R., van Rijsbergen, C.J.: The effectiveness of query-specific hierarchic clustering in information retrieval. *Information Processing & Management* 38(4), 559–582 (2002)
18. TREC, <http://trec.nist.gov/>
19. van Rijsbergen, C.J.: *Information Retrieval*. Butterworths (1979)
20. Wu, S., Bi, Y., McClean, S.: Regression relevance models for data fusion. In: *Proceedings of the 18th International Workshop on Database and Expert Systems Applications*, Regensburg, Germany, September 2007, pp. 264–268 (2007)
21. Wu, S., McClean, S.: Evaluation of system measures for incomplete relevance judgment in IR. In: Larsen, H.L., Pasi, G., Ortiz-Arroyo, D., Andreassen, T., Christiansen, H. (eds.) *FQAS 2006*. LNCS (LNAI), vol. 4027, pp. 245–256. Springer, Heidelberg (2006)
22. Zobel, J.: How reliable are the results of large-scale information retrieval experiments. In: *Proceedings of ACM SIGIR Conference*, Melbourne, Australia, August 1998, pp. 307–314 (1998)

# Autonomy: Life and Being

Mary-Anne Williams

Innovation and Enterprise Research Laboratory  
QCIS, University of Technology, Sydney  
Mary-Anne@TheMagicLab.org

**Abstract.** This paper uses robot experience to explore key concepts of autonomy, life and being. Unfortunately, there are no widely accepted definitions of autonomy, life or being. Using a new cognitive agent architecture we argue that autonomy is a key ingredient for both life and being, and set about exploring autonomy as a concept and a capability. Some schools of thought regard autonomy as the key characteristic that distinguishes a system from an agent; agents are systems with autonomy, but rarely is a definition of autonomy provided. Living entities are autonomous systems, and autonomy is vital to life. Intelligence presupposes autonomy too; what would it mean for a system to be intelligent but not exhibit any form of genuine autonomy. Our philosophical, scientific and legal understanding of autonomy and its implications is immature and as a result progress towards designing, building, managing, exploiting and regulating autonomous systems is retarded. In response we put forward a framework for exploring autonomy as a concept and capability based on a new cognitive architecture. Using this architecture tools and benchmarks can be developed to analyze and study autonomy in its own right as a means to further our understanding of autonomous systems, life and being. This endeavor would lead to important practical benefits for autonomous systems design and help determine the legal status of autonomous systems. It is only with a new enabling understanding of autonomy that the dream of Artificial Intelligence and Artificial Life can be realized. We argue that designing systems with genuine autonomy capabilities can be achieved by focusing on agent experiences of being rather than attempting to encode human experiences as symbolic knowledge and know-how in the artificial agents we build.

**Keywords:** Autonomy, agents, cognitive robotics, design.

## 1 Introduction

Machines like wristwatches, aircraft, smart phones and robots do things; they perform symbolic operations and undertake physical actions. Smart machines like mobile robots exhibit sophisticated behaviors; they perceive, recognize, and can move under the power of their own volition. However, they are not considered to be alive and we argue that is because they lack an essential ingredient, namely, genuine autonomy.

There is a discernable trend in the design and construction of machines with increasingly sophisticated capabilities that offer valuable intelligent services. The

introduction of computers has accelerated this trend towards machines that can conduct rapid computation and perform high precision actions. As machines become increasingly less dependent on human control they expose significant philosophical, scientific, technical and design challenges, as well as major social, ethical and legal conundrums.

Autonomy is an important and desirable trait for systems; it empowers them to enact decisions and pursue goals without human intervention. By way of illustration trends in advanced vehicle technology provide a compelling example of why we need to understand autonomy better. Automated technologies are becoming increasingly prevalent in the cars we drive and interact with on a daily basis. Consider power steering technology, cruise control, adaptive braking, autonomous parking systems, and pedestrian avoidance. The trend is towards more autonomous devices and more sophisticated autonomy; however there is no framework for autonomy, no measures of how autonomous a given system is relative to others. *Degree of autonomy* is judged only at the intuitive level where each assessor uses an entirely different set of criteria and measures.

When we interact with an autonomous system we give up certain aspects of control because such systems can act without our direct instruction and respond to things we cannot perceive. In order to work in partnership with autonomous systems we must have a clear understanding of the control we relinquish and of the capabilities of the system so that we can get the most value out of our interaction and deal with problems or system failures as they inevitably arise. In order to drive a car competently drivers need to know enough about the autonomous mechanisms in their car to know how to use them and about the circumstances under which they should seize back control and override the system. As a driver it is also important to know the consequences of overriding in a wide range of driving conditions as car control and engineering systems become increasingly autonomous. There are many spectacular examples<sup>1</sup> of real events where drivers have endured dramatic and terrifying experiences when autonomous technologies fail.

In this paper we seek to make a contribution by presenting a fresh approach to autonomy and its implications. We use an innovative cognitive framework based on representations of agent experiences to help explore the concept of autonomy and the underlying capabilities that support it. The framework is based on the idea that an agent needs to ground its own representations in its own experiences, to develop its own sense of being, so that it can motivate itself, control itself, learn for itself, act for itself and adapt to change. A successful intelligent autonomous agent must be able to make sense of unexpected and novel experiences, and to respond to them safely and appropriately.

## 2 Life and Autonomy

Autonomy, like life, is a complex concept that is difficult to define. We can recognize autonomy when we see it, but it is difficult to definitively characterize it in the form of a definition, or using set of features or properties.

---

<sup>1</sup><http://theage.drive.com.au/motor-news/cruise-control-terror-dramatic-triple0-tape-released-20091215-kuk8.html>

The influential philosopher Kant considered autonomy to be nonmechanistic and to be the embodiment of purpose. He argued that purposeful behavior is a key element of autonomy, however others consider purposefulness to be a property that an outside observer attributes to an autonomous system, rather than an intrinsic property of the systems itself. Living things appear to have a purpose, e.g. to stay alive.

Arguably the most autonomous systems currently in existence are biological entities; notwithstanding the debate around viruses, living biological autonomous agents essentially range in complexity from single cells to complex organisms like people. Consequently we have a rich source of autonomous systems to study, and model in artificial autonomous systems we design and the robots we build.

Plants and animals across the biological spectrum exhibit different cognitive capabilities and behaviors, which we take into consideration when we interact with them, e.g. apes are able to learn/modify behaviors that reptiles cannot, plants do not move but instead are phototropic and grow towards light.

Autonomy and artificial life are not only explored by computer scientists and AI researchers, but also by biologists, chemists, philosophers and others. For many autonomy is about self-organization and emergent behavior in dynamic systems. Thompson [18] identifies a spectrum of systems from *heteronomous* to *autonomous*. According to him heteronomous systems are other-governed; such system organization is defined by input-output information flow and external mechanisms of control. An autonomous system is self-governed; it has endogenous, self-organizing and self-controlling dynamics.

There is no definition of life. Biologists and philosophers tend to define and describe life using sets of characteristics like growth, adaptability, replication, and reproduction. Reproduction turns out to be a trait that is universally held by biologists and philosophers as crucial to life, however we find this entirely unsatisfactory when applied to entities like mobile robots which are more like hybrid than a species. The classic example of a hybrid is a mute, half horse and half donkey, and incapable of procreation. Many discussions., particularly those that hold reproduction dear, tend to focus on the *biology of life* rather than cognitive aspects of life; cognitive experience.

A key feature of existing artificial autonomous systems is that they are designed, engineered and programmed by people, and are not *physically or cognitively independent* in the same way that animals are independent.

The current approach to describing autonomy in the field of computer science, robotics and artificial intelligence tends to focus on the degree to which people control decision-making often with respect to action selection, e.g. transactions in software agents and behaviors in embodied agents like robots. We argue that genuine autonomy involves freedom in all cognitive areas including perception, conception, recognition, intention, motivation, awareness, attention, explanation, prediction and anticipation.

Traditionally the agent and robotics research communities, have focused on building systems that are grounded by human designers who encode their experience rather than by the system based on its experiences, thus producing systems that focus on achieving set tasks such as delivering the mail in an office scenario using human designed data structures for describing world states e.g. maps, rather than developing cognitive capabilities for an agent that would allow it to make sense of its own experiences. As a result agents, particularly robots, are incapable of interpreting their

own experiences and do not know what they are doing and why. Consequently, they typically perform poorly particularly in novel situations even when the amount of novelty they encounter is low.

The knowledge representation community has tended to focus on building models for artificial agents that can reason competently and deal with incompleteness, inconsistency, uncertainty, and change using symbolic logics. One of the major shortcomings of using these methods is that instead of providing a means for cognitive agents to develop their own representations, they encode knowledge and know-how based on human experience in high level logical languages, and consequently the resulting agents are incapable of dealing with novel experiences. In other words, encoding human knowledge and know-how typically leads to agent failure in complex dynamic environments when situations arise that have not been anticipated and hence not designed for by the human designers. In limited and closed domains traditional intelligent systems design methods that endow a computer system with human knowledge and know-how can be highly effective, but they tend not to be successful in open dynamic environments. McCarthy [25] identified this problem as a lack of elaboration tolerance in logic based systems.

Instead of developing systems that encode human experience, autonomous agents should learn to make sense of their own experiences. Autonomous systems need to be cognitively independent from their designers and have the means to develop important characteristics of intelligence like intuition, insight, creativity, motivation, intention, and curiosity.

A wealth of agent architectures have been developed for reactive agents, deliberative agents, and hybrid agents, together with agent-oriented toolkits and infrastructure [28]. However, there is a glaring and disturbing absence of high-level architectures for practical cognitive agents that explicitly support the design and development of capabilities for autonomy. Furthermore, most existing architectures require the designer to carefully and completely craft agents' representations of perceptions and behaviors so that they can achieve designated goals. The underlying assumption is that all possible scenarios have been catered for in the design. This approach is sound only in closed relatively static worlds where all relevant knowledge and know-how can be specified and encoded.

As a result despite the advances agent-oriented approaches offer over traditional software development approaches, the agents produced are typically incapable of dealing with unexpected and unforeseen circumstances that the designer did not prepare them for, or that require *making sense* of novel experiences [20]. This is a significant and debilitating problem for agents and robots with complex sensors in open environments where robust capabilities that can deal with and anticipate change are really needed.

Central to autonomy is *perceptualization*; the process that allows a cognitive agent to make sense of its own experience and to respond and behave in ways that align with its motivation. Perceptualization is the process of making sense of experience in our context involves making representations [20] for the purpose of recognizing objects and events, developing awareness and using it to anticipate future developments and eventualities.

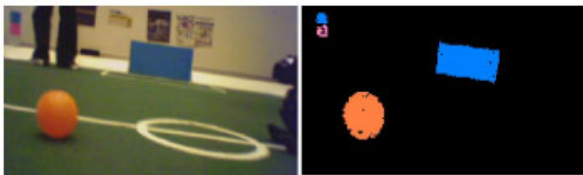
Agents can only know "reality" and give meaning to it through experience. Even the human body and mind is limited in its ability to determine the actual nature of

themselves and the world, and must construct representations of “reality” from experience.

When considering cognitive agent experiences, the adage that perception is reality rings true. For example, light in the real world is not colored and yet people experience colours, and believe that colours are a feature of the real world [12], similarly music is a construction of the mind. The way we experience or see color and hear music, and feel pain are examples of perceptualization; the construction of human experience representations.

It is important to note that representations are not grounded in the “real” world, but rather in experience. For example, there is no color in the external world to ground an internal representation; instead representations of color are grounded in our experience. Furthermore, there is no clear correspondence between color representations in our mind and the world since the color we experience is “constructed” from photons, ambient lighting conditions, the behavior of the 100 million rod and cone cells in our retina and other complex chemical reactions in the optic nerve and brain tissue. As Lakoff and Johnson [15] noted this very fact presents major obstacles to mainstream philosophy and AI, particularly symbolic logic approaches which are based on a correspondence theory of truth; because color provides a compelling counterexample that demonstrates there is no correspondence between a concept and the world, only between a concept and an experience. This obviously has a fundamental impact on current approaches to systems development, which tend to assume a correspondence between perception and the world not only exists but that it is perceptible or knowable.

How a cognitive agent represents and categorizes information depends on their (sensory, perceptual, and conceptual) experience, which in turn relies on specific cognitive capabilities and how they perceive, move and interact [15]. The kinds of representations an agent can create and share define it; human representation making defines us as human. Figure 1 highlights some differences between a person’s visual perception and a contemporary robot’s. If all a robot can experience and perceive via its camera are the pixel values or regions of pixels (so-called “blobs”) internally represented as strings of YUV/RGB values for each pixel or region data structures respectively, as visualized in Figure 1, then relative to a human its experiences are severely impoverished. The “representational distance” from an encoding of human experience to artificial agent experiences is substantive, but it is rarely taken into account in agent design. Instead agent designers develop clever, concise encodings, which are inscrutable from the actual agent’s perspective. Rather than building agent representations by encoding human-level hindsight, foresight and insight, we argue it is more important to develop cognitive capabilities and infrastructure for agents that would allow them to make sense of their own simple experiences all by themselves.



**Fig. 1.** Human vision verses robot vision experiences

### 3 Cognitive Independence

Autonomy is a multi-dimensional (morphology, control, motivation, understanding) and graded concept. In the realm of control, agents can vary from those that are remotely controlled by another agent to those that can autonomously develop representations to control all aspects of their existence all by themselves. We consider *cognitively independent* agents to be self-motivated, self-aware and self-governed, and able to make sense of their own experiences by building their own representations. A cognitively independent agent is fully autonomous. We believe that the kind of autonomy we should be striving for in the field of AI is cognitive independence.

People are obviously cognitively independent even though our behavior results from a physical and chemical based instructions encoded in our DNA.

A major challenge in the creation of cognitive independent robots is the development of representations of their experience; how best to enable a robot to conceptualize and describe them for its own internal purposes. If we have to design capabilities that allow a robot to create representations of its experience then we need a better understanding of experiences, how to describe and model them, how they are generated.

The cognitive experience architecture depicted in Figure 2 does not attempt to model the world directly or designers' perception and conception of it. Instead it supports representations of an agent's experience in the world directly. For example, robots have a morphology (a body), which is a physical mass, and in order to move it the agent must overcome a range of physical forces such as gravity and friction [19]. In addition to forces in the world there are relevant physical (and abstract) objects/things that the robot must perceive, recognize, attend and respond to through its bodily incarnation.

The cognitive experience architecture provides a useful tool to explore autonomy as a concept and as a capability. It has four main components grounded in the agent's experience: morphology, understanding, motivation and governance. The idea is to have an architecture that supports the design and development of agents that can make sense of their own experiences all by themselves guided by self generated motivations and cognitive capabilities for self-control and understanding.

In order for an agent to develop its own sense of being and self-determination, it requires core cognitive capabilities: morphology, motivation, governance and understanding. A self-determined agent must be able to represent and make sense of its own experience; use its motivations to drive control mechanisms and strategies; use their awareness and attention to drive understanding.

Agents like robots have a morphology; morphology can influence the autonomy of an agent. Body parts, degrees of freedom, sensors and actuators all play important roles in an agent's morphology and any framework purporting to describe autonomy must take morphology into consideration. Morphology determines an agent's sense of self, their self-concept [24, 29].

Agents experience themselves (via proprioception) and themselves in the world (perception). Proprioceptual and perceptual experience acquires information through internal and external sensors collectively.



Representations of experience are key to developing autonomous systems. Building mechanisms for agents to use to represent their experience rather than designing behaviors via human encoding of experience is the way forward.

Cognitive agents need to be self-motivated and to pursue their own goals such as seeking certain experiences depending on representations of needs, wants, current and future states. Motivation plays a crucial role in a cognitive agent and one must wonder what agency could possibly mean without it. The influence of an agent's motivation, intention and control mechanisms have a profound impact on autonomy.

Motivation can power agent control mechanisms and strategies. It influences what the agent is paying attention to, which in turn affects what the agent is currently aware of and how it might respond to its experiences. It determines what information will be grounded, how, when and why. People's attention and awareness are influenced by their passion, persistence, and perseverance all properties of motivation. At the very least autonomous robots require simple motivations: to act safely and not cause damage to themselves or others.

Understanding involves making sense of experience, which requires cognitive capabilities for making representations. *Representations possess affordances*; representations can afford action, behavior, reaction, deliberation, decision-making, learning, description, explanation, prediction, anticipation and many other capabilities.

Agents need to develop an awareness that can help them focus and inform attention both internally and externally.



**Fig. 2.** Cognitive Experience Architecture

The internal (e.g. body) and external world offers too much information to ground and subsequently process, therefore an agent must learn to select the information that will generate the most value. We know from neuroscience that prediction feedback

cycles are crucial; studies in neuropsychology suggest the value of comparing prediction information with perceived information, and in robotics prediction plays an important role and often implemented as Kalman and particle filters. There is an essential difference between prediction and anticipation. Prediction is forecasting the future, while anticipation involves prediction it also involves the crafting of appropriate responses to the prediction. Humans spend copious amounts of time thinking, rerunning old experiences and rehearsing future experiences as they anticipate and prepare for change. We need to understand these processes much better in order to develop genuinely intelligent autonomous systems.

Control mechanisms need to work in synergy with experience, regardless of how experience is modeled. The reason attention is so crucial is that it determines what an agent considers to be important or relevant right now for its current purpose which could be cues for what will happen next  $t$  at some stage in the future. Awareness and attention are intimately intertwined because awareness informs attention and vice-versa. A practical agent has bounded cognitive abilities and cannot give all incoming information due consideration. Instead it must focus on the most relevant things. Many animals, e.g. reptiles only recognize objects when they move because that is when they catch their attention. The key point is that autonomy is crucial to these cognitive activities.

Machine learning is a major and high successful area in AI, and it has led to significant breakthroughs and extraordinary insights, however there is still enormous scope to develop a new breed of high performance learning algorithms that an agent can utilize to learn quickly after experiencing a small number of examples [30]. Agents that can learn efficiently and iteratively all by themselves on the fly is a major AI challenge.

## 4 Social and Legal Challenges

Animals are good examples of autonomous agents and are often used to perform action and work. Our social and legal systems accommodate interactions between humans and autonomous systems like pets, farm animals, beasts of burden and sporting animals (polo, steeple-chasing, fox-hunting, horseracing). Society has developed expectations and laws around acceptable behavior of biological autonomous agents; lines of responsibility and liability are relatively clear and governed by law.

Working with animals can be hazardous, e.g. riding horses leads to more injuries than any other sport. There is an inherent risk working with biological agents largely due to their autonomy (rather than their biology) because the behavior of autonomous biological creatures can be difficult for people to anticipate in all circumstances because they experience the world differently to ourselves. Different circumstances may affect different animals in different ways at different times. Often people have to be trained to work with animals and the amount of training is typically proportional to the potential risks and skills required to control the animal. Furthermore, legal certification can also be required in some circumstances.

Dealing with autonomous systems is becoming a hot topic in legal circles as the recent panel on *Legal Challenges in the Age of Robots*<sup>2</sup> at the Stanford Law School attests. The ongoing debate struggles around the exact nature of autonomy with respect to a given autonomous system. Some open legal questions include:

1. How can relevant aspects of autonomy be described in ways that people can understand such that legal protection works in areas like safety, security and privacy?
2. How should liability be attributed to autonomous systems and their designers?
3. What laws are relevant to autonomous systems?
4. What laws are affected by autonomous systems?
5. What changes to the law are needed to accommodate the trends in autonomous system deployment?

Additional social and ethical questions that arise in view of the increasing sophistication of autonomous systems include:

1. If artificial autonomous systems have “recognized” experiences and can autonomously interpret them, what is their status in society?
2. What are the relevant dimensions of autonomy and how can we measure them?
3. What kinds of capabilities enable autonomy and what kinds of expected behaviors should they demonstrate to comply with society’s expectations?
4. How should people be prepared and trained to work with and interact with autonomous systems?
5. What kinds of cues and social mores do autonomous systems need to observe when they interact with humans and other agents?

## 5 Discussion

The lofty goals and ambitions of AI remain unchanged and unattained, however after more than 50 years of pursuing them we are well placed to take our understanding of autonomy to new levels. In this paper we have identified the need to develop new frameworks that allow us to build agents that can make their own representations and learn all by themselves.

The only way an agent can develop self-control and cognitive independence is if it has the capability to interpret and understand its own experiences. This requires the cognitive capability to build representations of their inner world experiences, and highlights the challenge of agile representation design.

Our discussion (re)raises deep philosophical issues and fuels the debate around the relationship between agents’ inner representations and the environment, and it challenges the prevailing system development paradigm. As usual there is no silver bullet however, the key challenge to developing genuine autonomy is determining how to endow agents with capabilities that will allow them to build effective representations of their own experiences.

---

<sup>2</sup> <http://www.law.stanford.edu/calendar/details/3496/>

In order to reach full autonomy robots will require the capability to understand and exploit their own experiences. Cognitive independence can be achieved by enhancement of morphological capabilities and developing methods for self-motivation, self-governance and understanding based on representations grounded in experience that an agent constructs all by itself. Clearly at this early stage cognitively independent robots will only be able to perform simple tasks because they are expected to achieve them all by themselves without human intervention.

Benchmarks of increasingly complexity designed to expose pertinent aspects of autonomy need to be developed. A community effort is required. Benchmarks can be used as tools for comparison, analysis and exploration of autonomy and its consequences.

After 50 years of AI research we still have a long way to go to achieve the original goal of building intelligent systems with human level perception, recognition and reasoning capabilities. Achieving the original goal of the field of AI was never going to be easy but unless more concerted effort is channeled into underlying fundamental problems, AI will never deliver its promise, make the breakthroughs needed to attain its ambitions or seize the prize it set out to grasp.

## References

1. Englemore, R., Morgan, A. (eds.): *Blackboard Systems*. Addison-Wesley, Reading (1986)
2. Anshar, M., Johnston, B., Novianto, R., Stanton, C., Wang, X., Williams, M.-A.: *The Bear Project: A Cognitive Approach to Robotics*. In: *ICAPS 2008, System Demonstration* (2008)
3. Baum, E.: *What is Thought*. MIT Press, Cambridge (2004)
4. Benferhat, S., Kaci, S., Le Berre, D., Williams, M.A.: *Weakening Conflicting Information for Iterated Revision and Knowledge Integration*. *Artificial Intelligence Journal* 153(1-2), 339–371 (2004)
5. Brooks, R.A.: *The Engineering of Physical Grounding*. In: *Proceedings of 15th Annual Meeting of the Cognitive Science Society*, pp. 153–154. Lawrence Erlbaum, Hillsdale (1993)
6. Brooks, R.A.: *The relationship between matter and life*. *Nature* (6818), 409–411 (2001)
7. Carnap, R.: *The logical structure of the World: Pseudoproblems in Philosophy*. Routledge, London (1967)
8. Hume, D.: *An Enquiry Concerning Human Understanding*. Hackett Publishing Company, Eric Steinberg (1977)
9. Ford, K.M., Glymour, C., Hayes, P.: *Thinking about Android Epistemology*. AAAI/MIT Press, Cambridge (2006)
10. Fujita, M.: *Intelligence Dynamics: A concept and preliminary experiments for open-ended learning agents*. *Journal of Autonomous Agents and Multi-Agent Systems* (to appear, 2009)
11. Gärdenfors, P., Williams, M.-A.: *Reasoning about Categories in Conceptual Spaces*. In: *Proceedings of the IJCAI. Morgan Kaufmann, San Francisco* (2001)
12. Goldstein, E.B.: *Sensation and Perception*, 6th edn. Wadsworth Publishing, Belmont (2002)
13. Johnston, B., Williams, M.-A.: *A formal framework for the symbol grounding problem*. In: *Proceedings of AGI* (2009)
14. King, J.R.: *Remaking the World: Modeling in Human Experience*. University of Illinois Press, Urbana (1996)

15. Lakoff, G., Johnson: *Philosophy in the Flesh: The Embodied Mind and Its Challenge to Western Thought*. Basic Books, New York (1999)
16. McCarthy, J.: The Well Designed Child. *Artificial Intelligence Journal* 172(18) (December 2008)
17. Steels, L.: Perceptually grounded meaning creation. In: *Proc of the International Conference on Multi-Agent Systems*. AAAI Press, Menlo Park
18. Thompson, E.: *Mind in Life*. Harvard University Press, Cambridge (2007)
19. Trieu, M., Williams, M.-A.: Grounded Representation Driven Robot Design. In: Lake-meyer, G., Sklar, E., Sorrenti, D.G., Takahashi, T. (eds.) *RoboCup 2006: Robot Soccer World Cup X*. LNCS (LNAI), vol. 4434, Springer, Heidelberg (2007)
20. Williams, M.-A.: Representation = Grounded Information. In: Wobcke, W., Zhang, M. (eds.) *AI 2008*. LNCS (LNAI), vol. 5360, pp. 42–48. Springer, Heidelberg (2008)
21. Williams, M.-A., McCarthy, J., Gärdenfors, P., Stanton, C., Karol, A.: A Grounding Framework. *Journal of Autonomous Agents and Multi-Agent Systems* 19, 272–296 (2009)
22. Varela, F., Thompson, E., Rosch, E.: *The Embodied Mind: Cognitive Science and Human Experience*. The MIT Press, Cambridge (1992)
23. Woolridge, M.: *An Introduction to MultiAgent Systems*. Wiley, Chichester (2009)
24. Novianto, R., Williams, M.-A.: The Role of Attention in Robot Self-Awareness. In: *The 18th IEEE International Symposium on Robot and Human Interactive Communication ROMAN 2009*, Toyama, September 2009. IEEE, Los Alamitos (2009)
25. McCarthy, J.: *Elaboration Tolerance* (2003),  
<http://www.formal.stanford.edu/jmc/elaboration/elaboration.html>
26. McCarthy, J.: *Notes on Self-Awareness*. In: *DARPA Workshop* (2004),  
<http://www-formal.stanford.edu/jmc/>
27. Gärdenfors, P.: *How Homo Became Sapiens: On the Evolution of Thinking*. Oxford University Press, Oxford (2003)
28. Luck, M., Ashri, R., d’Inverno, M.: *Agent-Based Software Development*, p. 208. Artech House, Boston (2004)
29. Pfeifer, R., Bongard, J.C.: *How the Body Shapes the Way We Think A New View of Intelligence*. MIT Press, Cambridge (2006)
30. Johnston, B., Williams, M.: *Autonomous Learning of Commonsense Simulations*. In: *International Symposium on Logical Formalizations of Commonsense Reasoning*, pp. 73–78 (2009)

# A Method of Social Collaboration and Knowledge Sharing Acceleration for e-Learning System: The Distance Learning Network Scenario

Przemysław Różewski

West Pomeranian University of Technology in Szczecin,  
Faculty of Computer Science and Information Systems,  
ul. Żołnierska 49, 71-210 Szczecin, Poland  
prozewski@wi.zut.edu.pl

**Abstract.** Nowadays, e-learning systems take the form of the Distance Learning Network (DLN) due to widespread use and accessibility of the Internet and networked e-learning services. The focal point of the DLN performance is efficiency of knowledge processing in asynchronous learning mode and facilitating cooperation between students. In addition, the DLN articulates attention to social aspects of the learning process as well. In this paper, a method for the DLN development is proposed. The main research objectives for the proposed method are the processes of acceleration of social collaboration and knowledge sharing in the DLN. The method introduces knowledge-disposed agents (who represent students in educational scenarios) that form a network of individuals aimed to increase their competence. For every agent the competence expansion process is formulated. Based on that outcome the process of dynamic network formation performed on the social and knowledge levels. The method utilizes formal apparatuses of competence set and network game theories combined with an agent system-based approach.

**Keywords:** Community-based learning, e-learning, competence management, social network, agents-based learning system.

## 1 Introduction

The modern educational market is polarized by the Open and Distance Learning (ODL) concept [26]. It assumes open and unrestricted access to didactical material and services using the Internet and e-learning technologies. At the same time the latter one allows to maintain a high level of security and efficiency of communication. Geographical dislocation of knowledge sources and students are reasons for evaluation of a e-learning system in the Distance Learning Network (DLN). In such network its nodes actively process knowledge and edges represent channels for knowledge relocation [2,4].

The important part of the DLN concept is a competence-based learning. Furthermore, supporting the concept of competence-based education and competence-based learning requires a search for proper recognition of the term competence that associates

the learning process with competences [22]. We can define competence as some ability to find an effective way of using theoretical knowledge to solve practical tasks and to verify the determined solution [23]. The foundation for competence is procedural knowledge associated with appropriate theoretical knowledge. In our approach the process of competence acquisition can be accelerated by maintaining a relationship with other students. Interestingly, some examples of competences can only be acquired exploiting such relationship with other students. We focus on the concept of competence basing on the following assumption: when a student possesses some competence it is indicative of one's ability to solve domain-oriented problems.

On the current e-learning market the asynchronous learning mode is becoming most important one [3]. In that mode a student has access to personalized didactical material at any time using any available locations. The postulation of asynchronous learning mode takes for granted the idea of a student learning without supervision merely supported by the DLN. Because of that, it is important to supply the DLN with social network capabilities [23]. The DLN should be able to connect students with similar activities, problems and tasks. More specifically in that sense it is essential to maintain emergence and synergy in the learning/teaching process as well as student group activation, student support, student workplace support and content support. What is more, efficient DLN network design, aside from social characteristics, supports knowledge and subject-related resource exchange. In addition, the DLN both supports and encourages individuals to learn together while retaining individual control over their time, space, presence, activity, identity and relationships.

The research problem addressed in the paper is formulated as creating a social group according to the criterion of domain interest similarity. The social group is sustained by the DLN (i.e., collaborative learning environment) and the DLN's objective is to associate learning individuals with similar and knowledge-based learning objectives in order to facilitate competence-based learning. In the framework of student's competence expansion process the competences are classified based on the competence set theory [33]. Afterwards, for every student a corresponding agent is created. The agent's cooperation environment is based on the network game theory [6,9]. The network game theory allows to model dynamic network formation. Information about student's based competence and learning goals are incorporated into the agents. Every agent interacts with peer agents and based on competence expansion cost method and social relations the connections between the agents are created. Each agent is poised to determine optimal relationship for overall minimal competence expansion cost and maximal social benefits.

The proposed approach is based on computational methods, in contrast to other soft-method approaches. The formal models allow to create information system for knowledge management in e-learning. The presented algorithms are founded on data structures and standards already known in the e-learning literature.

## **2 The Concept of the Distance Learning Network**

The reason behind the Distance Learning Network (DLN) development is to make whole spectrum of didactical materials and services accessible for students and

maintain their cooperation and social relationships. According to works of the UE TenCompetence Project [13,14,27] the DLN connects actors, both human and agents, with institutions and learning resources. In order to meet that requirement an advanced information and communication infrastructure is required. Moreover, a capable corporation network should connect all parts of the DLN. In addition, efficient work organization allows to proper role distribution between social agent and access to services and tools.

## 2.1 Knowledge Aspects of the Distance Learning Network

In the asynchronous learning mode, the structure of didactic materials is module-based. Every domain can be divided into modules (called Learning Objects) consistent with the SCORM e-learning standard that can be later used in different courses, without the need for expensive recoding or re-designing operations [34]. The didactic material acquires the features of a knowledge base and the teacher performs only activities such as consulting and student knowledge evaluation. This means that the teacher participates in courseware design as knowledge engineer and tests the students' absorbed knowledge [16]. Student works with didactical material on his/her own based on the network capabilities.

**Table 1.** Comparison of different analysis approaches to the Distance Learning Network

<i>Design concept</i>	<i>Telecommunications network</i>	<i>Workflow network</i>	<i>Agent's network (knowledge network)</i>
<i>Analytical basis</i>	Queuing theory	Graph theory	Network game theory, competence set theory, ontology
<i>Network's unit</i>	Packet	Task	Competence set, Knowledge/information object, concept
<i>Control parameters</i>	Security, speed	Efficiency, workload	Communication efficiency, Completeness, credibility
<i>Node concept</i>	Computer station	Work station	Social agent
<i>Node role</i>	Signal regeneration, data distribution	Technological operation	Competence, knowledge and information creation
<i>Work paradigm</i>	Standards	Technological chain	Emergency, synergy

Every node of DLN is a student, who can be represented by an agent (social agent). The agent's objective is to create new knowledge according to his/her logic in order to develop a certain competence. Additionally, every agent is characterized by a corresponding competence set. The reasons for knowledge creation are dependent on the learning/teaching objectives specific to every DLN agent. According to the knowledge flow concept [37] agents (network node) plays a role of knowledge portal or knowledge process with different intensity. The node can either generate, learn,



process, understand, synthesize, and deliver knowledge [36]. Complete integration of different knowledge sources is possible based on e-learning standards e.g., SCORM, IEEE LOM, CORDRA [16,34].

The complex nature of the DLN can be analyzed on many levels (see Table 1). The presented discussion will focus on the agent network aspect. On that level social communication allows to increase student's competence due to its knowledge-based nature.

## 2.2 Social Aspects of the Distance Learning Network

The collaboration process is the most important social activity occurring in the DLN. During the collaboration the individuals work jointly over certain portion of knowledge. Collaboration facilitates acquisition and integration of knowledge resources through external integration and cooperation with other cooperative or supporting agents conducted on a basis of common consensus, trust, cooperation, and sharing by a multi-functional team of experienced knowledge workers (students). From the learning process point of view, the collaboration requires consensus, mutual understanding, reciprocity and trust [25]. From the knowledge sharing point of view, the collaboration requires complete understanding and effective sharing of information and knowledge throughout the development cycle [8].

The node in the DLN plays social role. Every node can play more than one role. According to e-quality project [7] the learning-teaching process is divided into following processes (including related roles): Planning process (instructional designer, content planner), Administration process (educational administrator, coordinator, technical administrator, advisor), Evaluation process (students' evaluator, developer), Learning material production process (material designer, material producer, audio-visual specialist) and Student's support process (pedagogical support, technological support, tutor, teacher, student).

## 3 Related Work

In the e-learning process the concept of social network is applied on the cognitive, knowledge and social interaction levels (see Table 2). Moreover, the social network concept extends the Learning Management System with educational social software [2] i.e., networked tools that support and encourage individuals to learn together while retaining individual control over their time, space, presence, activity, identity and relationship.

The vital element of the discussed research problem comes from the idea of social network development using the agent approach (but limited to distance learning systems). Due to its high flexibility, the MAS's approach is popular in distance learning system. The agent's approach is usually applied for distance learning system development [36]. The MAS's approach is used on the system modules level. On the contrary in discussed approach the MAS is used for relations between students and resources development in the working DLN conditions.

**Table 2.** Different social network aspects of e-learning systems

Idea	Focal point	References
Relations between social network characteristics in an online class and <b>cognitive learning outcomes</b>	Collaborative Learning environment, sense of community, prestige	[4,19,20,24]
Social network analysis provides meaningful and quantitative insights into the quality of <b>knowledge construction process</b>	Knowledge process optimization, knowledge flow,	[2,5,17,32]
<b>Interactions in social groups</b> and the impact of intrinsic characteristics of individuals on their <b>social interactions</b> .	Social architectures	[28,30]

## 4 The Concept of Distance Learning Network Development

The design objective is to develop a DLN creation method that supports social collaboration and knowledge sharing processes for the e-learning asynchronous learning mode environment. For that propose the competence expansion process is discussed. During that process the student acquires new competence(s). The competences are grouped by its respective types. In cases of a “difficult” competence the collaboration with other students is set in the framework of the dynamic network formation process. The agents (students) connected with other agents possess knowledge (in form of competences) that directly assists ongoing competence acquisition process within the network. Moreover, the student’s operational environment is responsible for his/her intelligent stimulation. The social context is accelerated by community of collaborators, distributed cognition, and learning community. In addition, social collaboration on the cognitive level is supported by joint formation of knowledge, and reciprocal sense making.

### 4.1 Theoretical Basis

In the proposed approach two different theories have been combined: network game theory and competence set theory. The network game theory allows to describe the agent-to-agent interaction on the social and knowledge levels while the competence set theory is used to explain agent competence expansion process.

We can recognize two approaches to the DLN (instance of social network) analysis. The first approach is focused on the model development of the network. Typical research problems are related to either some stochastic process or an algorithmic through which the links in the network are formed [1]. The second approach, called network game theory, is mainly concerned with models where the links are formed at the discretion of the nodes who derive benefits and face costs associated with various links and network configurations [10]. That approach is used in the paper. The network game theory was developed as an approach to formulate networks connecting individuals. As it has been showed in [6], the organization of individual agents into networks and groups has an important role in the determination of the outcome of many social and economic interactions.

The students (social agents) activities in the DLN can be described in terms of network game theory. This theory assumes that payoff to agents from an economic or social activity depends on the network of connections among agents [12]. Methods of keeping track of the overall value generated by a particular network, as well as how it is allocated across players (students – agents), are through a value function and an allocation rule [11].

In order to calculate the value function and a rule for allocating every student in the network it is important to estimate the potential benefits for students appearing in different network configurations. Based on the competence set theory it is possible to calculate the student-specific benefit from the actual configuration of the network. The relationship between the students with well-matched knowledge allows to maintain collaboration process, which is the firmament of social network. The competence set is defined as the set containing skills, information and knowledge of any agent [31]. What is more, the competence expansion cost can be calculated in relation to any agents configuration in the DLN [15]. Students with similar competence can achieve similar learning goal equally attaining low cost.

## 4.2 Competence Expansion Process

The student possesses some set of competence form his/her previous activities. We can recognize this competence set as background knowledge (acquired competence set  $Sk$ ). The competence set related to learning goals (required competence set  $Tr$ ) includes all competences with corresponding strength level which student has to obtained. This process is described as obtaining new competences and adding them to the actual acquired competence set of a person. The cost and pace of obtaining new competences depends on elements of current competence set and how close these elements are related with the new competences. Theoretical basis for competence assessment can be found in [18].

Let's introduce some formal notions based on [15,29]. For each competence  $c$ , its fuzzy strength  $\alpha: \{s\} \rightarrow [0;1]$  is a function of a student  $s_i$  in the context of which the competence is assessed. The learning goals are made up of the set of competences  $E \in \{c_j\}$ . If competence  $c_1$  is a background competence of competence  $c_2$ , there exists a background relation denoted by  $r(c_1, c_2)$  between these two competences  $0 < r(c_1, c_2) = r_{21} \leq 1$ . If  $r_{21} = 0$  then competence  $c_2$  is not a background competence of  $c_1$ . For any student  $S$ , if competence  $c_1$  with strength  $\alpha_1$  is the only background competence of  $c_2$ , then the background-strength (potential) of obtaining  $c_2$  equals  $\beta_2 = \alpha_1 \cdot r_{21}$ . If a competence has several background competences then the learning process progresses according to the principle of maximal support [29]. For the student  $s$ , facing learning goal  $E$  requiring competence  $c$  with strength  $\alpha$  (denoted by  $c^\alpha(E)$ ), the critical level  $\gamma$  of the background strength  $\beta$  needed to obtain competence  $c$  is the value from range  $[0; 1]$ , so that if  $\beta \geq \gamma$  then  $Cost = 0$ , otherwise  $Cost > 0$ . The value of the critical level  $\gamma$  depends on the competence set and is denoted by  $\gamma(c)$ . A student

willing to obtain competence  $c$  has to possess such background strength of this competence, that  $\beta \geq \gamma(c)$ . If the condition is met, then competence  $c$  is called the person's skill competence  $Sk(s)$ . Otherwise,  $c$  is called the person's non-skill competence  $NSk(s)$ .

We can classify student  $s_i$  competences with relation to learning goals ( $E$ ) as follows [15,29]:

- 1) If  $\exists_{i=1}^n (\beta(s_i) \geq \beta(E) \wedge \alpha(s_i) \geq \alpha(E))$ , then competence  $c^{\alpha,\beta}(E)$  is called a type (1) group competence;
- 2) If  $c^{\alpha,\beta}(E)$  is not a type (1) competence and  $\exists_{i=1}^n (\beta(s_i) \geq \beta(E) \wedge \alpha(s_i) < \alpha(E))$ , then this competence is a type (2) competence
- 3) If  $\forall_{i=1}^n \beta(s_i) < \beta(E)$ , then competence  $c^{\alpha,\beta}(E)$  is called a competence type (3)

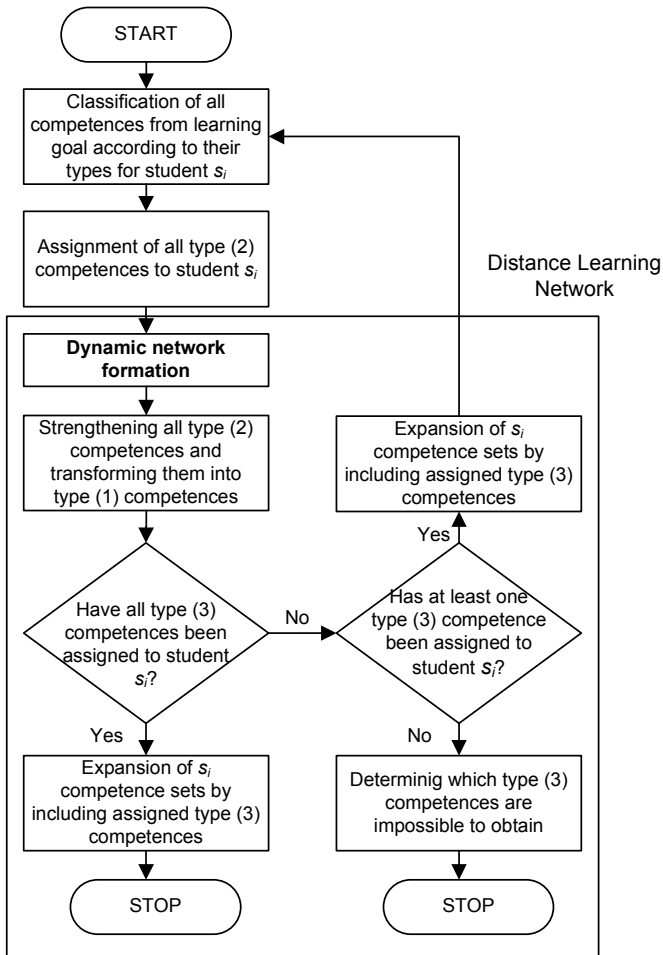


Fig. 1. The competence expansion algorithm

The competence classification can be used for students cooperation planning development (based on the agents network). On the Figure 1 the student competence expansion algorithm is presented. The algorithm results are: (i) initial information for agent, (ii) cost assumption for each competence expansion process, (iii) competences expansion process sequence. In the first step of algorithm all competence related with learning goal ( $E$ ) are classified according to their respective types. When the classified competences are of type (1) then the student possesses all required competences with sufficient level of background strength to achieve learning goals. In case of competence of type (2) the student can strengthen (at some additional cost) all competences of type (2) from level  $\alpha(s_i)$  to required level  $\beta(E)$ . The cost of the strengthening process is narrowed by the cooperation with other students in the framework of agent's network DLN. In case of competence of type (3) the student cannot strengthen all competences type (3) because of all this competences do not belong to the student's  $Sk$  set. The student has to increase his/her background strength  $\beta(s_i)$  through cooperation with other students and than strengthening competence's strength.

The critical element of the competence expansion algorithm is a dynamic network formation process. The cooperation with other students in the DLN framework is facilitated with the achievement of competences of types (2) and (3). During the dynamic network formation process an agent is created, with the purpose of finding best students to cooperate with. The criterion of best competence suitability, regarding to cost, is applied. The dynamic network formation process is based on the network game theory presented in the next section.

### 4.3 Dynamic Network Formation

The network relations among agents, who represent students, are formally represented by undirected graph whose nodes are identified with the agents and whose arcs capture the pairwise relations [6,9]. The utility function  $u_i$  is a function that assigns a payoff to every agent as a function of the underlying network  $g$  connecting them. The value function allows the value of a network to depend in arbitrary ways on the structure of the networks. The different networks that connect the same agents may lead to different values.

Beyond knowing how much total value is generated by a network, it is critical to keep track of how that value is allocated or distributed among the agents in the DLN. This is captured by the concept of the allocation rule. An allocation rule describes how the value associated with each network is distributed to the individual agents.  $Y_i(g, u)$  is the payoff to agent  $i$  from graph  $g$  under the value function  $u_i$ .

The allocation rule, in the context of the DLN, should incorporate costs of individual links as well as student benefits. The agent's cost comes from the time and resources spent to maintain relations with other students. In that sense typical activities are: consultation, comments and discussion. The benefits come from intellectual collaboration with students with similar education level (i.e. similar learning goal or/and background knowledge). Moreover, the student cannot achieve the type (3) competence on his/her own without collaboration with other students. The creation of relation between the students is important from the point of view of collaborative

learning environment. The DLN consists of several sub-networks which create social network dedicated to specific task/concept/competence. In the dynamic network formation process special attention should be paid to most efficient student's connection (optimal DLN structure). For student  $s1$  and  $s2$  the following interaction between competences and learning goals may happens:

- Synergy: students  $s1$  and  $s2$  have similar learning goals and background competence with sufficient strength. For each part such relation generates best profit with lower cost.
- Emergence: students  $s1$  and  $s2$  have different learning goals, however background competences are adequate. Because the students' learning goals are different one time cooperation are possible only in case of single competence.
- Cooperation: students  $s1$  and  $s2$  have similar learning goals and inadequate background competences. The students are aiming to the same goal. The difference in background competences may cause some cost related to student's knowledge adaptation.
- Non-relation: there is no intersecting between students learning goals and background competences. The cooperation in such case is usually pointless.

Additionally, the number of relations between agents affects the agent relation quality. Greater numbers of relations increase the cost of relation, because of the student's recourses must be used more efficiently to ensure proper relation quality. The agents allocation rule should incorporate interaction between students' competences and learning goals as well as the number of agent relations. Moreover, some results of cognitive and pedagogical sciences may also be used. A number of models can be found in network game theory literature [6,9]. However the models of allocation rule especially for the DLN must be developed separately.

The dynamic network formation process runs continuously. Every agent can disconnect from the DLN due to accomplishment of the learning goals or allocation rules redefinition.

## 5 An Implementation Issue of Proposed Method

Implementation of the presented method requires both the computer support and didactic method interpretation. The computer support supplies tools and systems for creating a social group according to the criterion of domain interest similarity based on the processing of competences}. The didactic method interpretation is required to determined proper e-learning context and suitable learning activities.

The computer implementation of the social collaboration and knowledge sharing acceleration method necessitates development of the information system's competence processing components. The proposed competence processing method is based on the competence set theory [33] united with TENCompetence, a competence modeling method [27], and the HR-XML and IEEE RCD standards. All those approaches have been combined in the Competence Object Library (COL) [22]. The COL standard deals with overall competence structure modeling as well as the computational mechanism of the competence expansion modelling. The COL integrates the fuzzy

competence set approach and method of analysis of competence expansion cost. In case of the discussed method the COL's methods: `CompetenceSet.CompareSet()` and `Competence-Profile.CompareProfile()` are particularly helpful. The computer infrastructure is the next element of the real life implementation in the discussed method. The proposed infrastructure can be found in [21] in the form of system for competence-based individual project selection and collaboration support. The infrastructure allows to store and process a competence object. In the discussed system the information about competences are collected from the following outcomes: (i) enrolment analysis, (ii) continuous learning progress evaluation, (iii) individual work evaluation. The competences are stored in the personal competence database in a form of personal competence profiles. In the proposed system, the student can be represented by an agent, whose communication language is extended by the COL [21].

Since we have taken the advantage of the competence concept, an advanced knowledge modelling in the different e-learning scenario is possible. For example, the proposed method can be used in e-learning environment for facilitating and increase of the creativity factor [21]. A creative distance learning environment allows to match task (or project) with student's creativity potential. In this case both the value function and the allocation rule are related to the student's creativity potential, his/her knowledge and knowledge of other students.

## 6 Conclusions

Modern distance learning system is networked and we should thrive to take advantage of that situation. Apart from the research issues of distributed resources and work organization it is reasonable to conclude that one should put attention to network social and knowledge aspects of the learning process. The proper DLN organization requires not only an approach to agent implementation but also an approach to agent-based learning environment development. Such approach should allow to evaluate cost and benefit of different DLN configurations. In that context, application of network game theory not only does it allow to make such forecast but also creates mechanism for agent interactions. In order to create a complete solution, a more detailed model of agent allocation is required.

The social collaboration is supported by a student relationship with other students (connections between agents). The proposed method promotes the individuals with most domain-oriented connections. On the knowledge level, the connected students share common knowledge (i.e. in form of an ontology). That situation increases the efficiency of communication and mutual understanding. In such case knowledge is shared during social learning process, which is maintained by the discussed method.

## References

1. Albert, R., Barabási, A.-L.: Statistical mechanics of complex networks. *Reviews of Modern Physics* 74, 47–97 (2002)
2. Ben, C., Nien-Heng, C., Yi-Chan, D., Tak-Wai, C.: Environmental design for a structured network learning society. *Computer & Education* 48(2), 234–249 (2007)
3. Bourne, J., Harris, D., Mayadas, F.: Online Engineering Education: Learning Anywhere, Anytime. *Journal of Engineering Education* 94(1), 131–146 (2005)

4. Cho, H., Gay, G., Davidson, B., Ingraffea, A.: Social networks, communication styles, and learning performance in a CSCL community. *Computer & Education* 49(2), 309–329 (2007)
5. Daniel, B., McCalla, G., Schwier, R.: Social Network Analysis techniques: implications for information and knowledge sharing in virtual learning communities. *International Journal of Advanced Media and Communication* 2(2), 20–34 (2008)
6. Dutta, B.D., Jackson, M.O.: *Networks and Groups: Models of Strategic Formation*. Studies in Economic Design. Springer, Heidelberg (2003)
7. e-Quality: Quality implementation in open and distance learning in a multicultural European environment, the Socrates/Minerva European Union Project (2003–2006), <http://www.e-quality-heu.org/>
8. Ho, C.-T., Chen, Y.-M., Chen, Y.-J., Wang, C.-B.: Developing a distributed knowledge model for knowledge management in collaborative development and implementation of an enterprise system. *Robotics and Computer-Integrated Manufacturing* 20(5), 439–456 (2004)
9. Jackson, M.O.: *Social and Economic Networks*. Princeton University Press, Princeton (2008)
10. Jackson, M.O.: *The Economics of Social Networks*. In: Blundell, R., Newey, W., Persson, T. (eds.) *Ninth World Congress of the Econometric Society. Advances in Economics and Econometrics, Theory and Applications*, vol. I. Cambridge University Press, Cambridge (2006)
11. Jackson, M.O.: *A Survey of Models of Network Formation: Stability and Efficiency*. In: Demange, G., Wooders, M. (eds.) *Group Formation in Economics; Networks, Clubs and Coalitions*, pp. 11–57. Cambridge University Press, Cambridge (2004)
12. Jackson, M.O., Watts, A.: *The Evolution of Social And Economic Networks*. *Journal of Economic Theory* 106(2), 265–295 (2002)
13. Kalz, M., Van Bruggen, J., Rusmann, E., Giesbers, B., Koper, R.: Positioning of Learners in Learning Networks with Content-Analysis, Metadata and Ontologies. *Interactive Learning Environments* 15, 191–200 (2007)
14. Koper, R., Rusman, E., Sloep, P.: connecting people, organisations, software agents and learning resources to establish the emergence of effective lifelong learning. *LLine: Lifelong Learning in Europe* 9(1), 18–27 (2005)
15. Kusztina, E., Zaikin, O., Rózewski, P., Małachowski, B.: Cost estimation algorithm and decision-making model for curriculum modification in educational organization. *European Journal of Operational Research* 197(2), 752–763 (2009)
16. Kushtina, E., Zaikin, O., Rózewski, P.: On the knowledge repository design and management in E-Learning. In: Lu, J., Ruan, D., Zhang, G. (eds.) *E-Service Intelligence: Methodologies, Technologies and applications*. Studies in Computational Intelligence, vol. 37, pp. 497–517. Springer, Heidelberg (2007)
17. Liberman, S., Wolf, K.B.: The flow of knowledge: Scientific contacts in formal meetings. *Social Networks* 19(3), 271–283 (1997)
18. Massaro, D.W.: A computer-animated tutor for language learning: Research and applications. In: Spencer, P.E., Marshark, M. (eds.) *Advances in the spoken language development of deaf and hard-of-hearing children*, pp. 212–243. Oxford University Press, New York (2006)
19. Palonen, T., Hakkarainen, K.: Patterns of Interaction in Computer-supported Learning: A Social Network Analysis. In: Fishman, B., O'Connor-Divellbiss, S. (eds.) *Proceedings of the Fourth International Conference of the Learning Sciences*, pp. 334–339. Erlbaum, Mahwah (2000)



20. Russo, T.C., Koesten, J.: Prestige, centrality, and learning: a social network analysis of an online class. *Communication Education* 54(3), 251–254 (2005)
21. Rózewski, P., Małachowski, B.: System For Creative Distance Learning Environment Development Based On Competence Management. In: KES 2010. LNCS (LNAI), Springer, Heidelberg (2010) (accepted, in press)
22. Rózewski, P., Małachowski, B.: Competence Management In Knowledge-Based Organisation: Case Study Based On Higher Education Organisation. In: Goebel, R., Siekmann, J., Wahlster, W. (eds.) KSEM 2009. LNCS, vol. 5914, pp. 358–369. Springer, Heidelberg (2009)
23. Rózewski, P., Ciszczyk, M.: Model of a collaboration environment for knowledge management in competence based learning. In: Nguyen, N.T., Kowalczyk, R., Chen, S.-M. (eds.) ICCCI 2009. LNCS (LNAI), vol. 5796, pp. 333–344. Springer, Heidelberg (2009)
24. Shen, D., Nuankhieo, P., Huang, X., Amelung, C., Laffey, J.: Using Social Network Analysis to Understand Sense of Community in an Online Learning Environment. *Journal of Educational Computing Research* 39(1), 17–36 (2008)
25. Skyrme, D.J.: The realities of virtuality. In: Sieber, P., Griese, J. (eds.) *Organizational Virtualness Proceedings of the VO Net – Workshop*, April 1998. Simowa Verlag, Bern (1998)
26. Tait, A.: Open and Distance Learning Policy in the European Union 1985-1995. *Higher Education Policy* 9(3), 221–238 (1996)
27. TENCompetence - Building the European Network for Lifelong Competence Development (2005-2009), EU IST-TEL project, <http://www.tencompetence.org/>
28. Trier, M., Bobrik, A.: Social Search: Exploring and Searching Social Architectures in Digital Networks. *IEEE Internet Computing* 13(2), 51–59 (2009)
29. Wang, H.-F., Wang, C.H.: Modeling of optimal expansion of a fuzzy competence set. *International Transactions in Operational Research* 5(5), 413–424 (1995)
30. Warren, A.R.: Network Analysis of Social Interactions in Laboratories. In: 2008 Physics Education Research Conference, AIP Conference Proceedings, Edmonton, vol. 1064, pp. 219–222 (2008)
31. Hu, Y.-C., Chen, R.-S., Tzeng, G.-H.: Generating learning sequences for decision makers through data mining and competence set expansion. *IEEE Transactions on Systems, Man, and Cybernetics, Part B* 32(5), 679–686 (2002)
32. Yonggu, W., Xiaojuan, L.: Social Network Analysis of Interaction in Online Learning Communities. In: Seventh IEEE International Conference on Advanced Learning Technologies, ICALT 2007, pp. 699–700. IEEE Computer Society, Niigata (2007)
33. Yu, P.L., Zhang, D.: A foundation for competence set analysis. *Mathematical Social Sciences* 20, 251–299 (1990)
34. Zaikin, O., Kushtina, E., Rózewski, P.: Model and algorithm of the conceptual scheme formation for knowledge domain in distance learning. *European Journal of Operational Research* 175(3), 1379–1399 (2006)
35. Liu, Z., Chen, B.: Model and Implement an Agent Oriented E-Learning System. In: International Conference on Computational Intelligence for Modelling, Control and Automation, 2005 and International Conference on Intelligent Agents, Web Technologies and Internet Commerce, vol. 2, pp. 859–864 (2005)
36. Zhuge, H.: Knowledge flow network planning and simulation. *Decision Support Systems* 42(2), 571–592 (2006)
37. Zhuge, H.: Discovery of Knowledge Flow in Science. *Communications of the ACM* 49(5), 101–107 (2006)

# A Comparative Study of Target-Based Evaluation of Traditional Craft Patterns Using Kansei Data

Van-Nam Huynh, Yoshiteru Nakamori, and Hongbin Yan

School of Knowledge Science  
Japan Advanced Institute of Science and Technology  
Nomi, Ishikawa, 923-1292, Japan  
huynh@jaist.ac.jp

**Abstract.** Evaluation for ranking is very useful for users in their decision-making process when they want to select some item(s) from a large number of items using their personal preferences. In this paper, we will focus on the evaluation of Japanese traditional crafts, in which product items are assessed according to the so-called *Kansei* features by means of the semantic differential method. In particular, two decision analysis based evaluation procedures, which take consumer-specified preferences on kansei features of traditional products into consideration, will be discussed and compared.

## 1 Introduction

Nowadays, in an increasingly competitive world market, it is important for manufacturers to have a customer-focused approach in order to improve attractiveness in development of new products, which should satisfy not only requirements of physical quality, defined objectively, but also consumers' psychological needs, by essence subjective [9]. This approach has actually received much attention since the 1970s from the research community of consumer-focused design and Kansei engineering, which is defined as "translating technology of a consumer's feeling and image for a product into design elements" [7]. Kansei engineering has been developed and successfully applied to a variety of industries, especially, in Japan. *Kansei* is a Japanese term which, according to Mitsuo Nagamachi – the founder of Kansei engineering, is 'the impression somebody gets from a certain artefact, environment or situation using all her senses of sight, hearing, feeling, smell, taste [and sense of balance] as well as their recognition' as quoted from [11]. For building a kansei database on psychological feelings regarding products, the most commonly-used method is to choose (adjectival) kansei words first, and then ask people to express their feelings using those kansei words by means of the semantic differential (SD) method [8].

The focus of this paper is on the evaluation of traditional craft products for personalized recommendation using kansei data, taking consumer-specified preferences on kansei features of traditional products into consideration. It should

be emphasized here that artistic and aesthetic aspects play a crucial role in perception of traditional crafts, therefore kansei data are essential and necessary for evaluation. Such evaluation would be helpful for marketing or personalized recommendation, which is particularly important in the current service-oriented economy where recommender systems are gaining widespread acceptance in e-commerce applications [1,3]. In [6], we have developed a consumer-oriented evaluation model for traditional Japanese crafts based on the appealing idea of target-based decision analysis [2]. Particularly, given a consumer's request, using available kansei assessment data the developed model aims to define an evaluation function that quantifies how well a product item meets the consumer's feeling preferences.

Recently, Martínez [12] has proposed to use linguistic decision analysis for sensory evaluation based on the linguistic 2-tuple representation model [4]. Note that the knowledge used for sensory evaluation is also acquired by means of human senses of *sight*, *taste*, *touch*, *smell* and *hearing*. Basically, Martínez's model considers the evaluation problem as a multi-expert/multi-criteria decision-making problem and assumes a consistent order relation over the qualitative evaluation scale treated as linguistic term set of a linguistic variable [16]. In fact, Martínez's model yields an overall ranking of evaluated objects, which is therefore inappropriate for the purpose of personalized recommendations.

In this paper we will first customize the linguistic 2-tuple representation model to make it applicable to the consumer-oriented evaluation problem for traditional Japanese crafts using kansei data, and then conduct a comparative study of these two methods. The rest of this paper is organized as follows. Section 2 describes the consumer-oriented evaluation problem using kansei data for traditional crafts. Section 3 introduces two decision analysis based methods for solving the consumer-oriented evaluation problem, one is based on fuzzy target-based decision analysis and the other is based on the linguistic decision analysis using the 2-tuple linguistic representation model. Section 4 then provides an illustration of these methods to evaluation of Kutani coffee cups along with a comparative analysis of the obtained results. Finally, some conclusions are presented in Section 5.

## 2 Kansei-Based Evaluation Problem

For traditional crafts, decisions on which items to buy or use are usually influenced by personal feelings/characteristics, so an evaluation targeting those specific requests by consumers would be very useful, particularly for the purpose of personalized recommendation. In this section, we will describe such a consumer-oriented evaluation problem using kansei data for traditional crafts [13]. Let us denote  $\mathcal{O}$  the collection of craft patterns to be evaluated and  $N$  is the cardinality of  $\mathcal{O}$ , i.e.  $N = |\mathcal{O}|$ .

The first task in the Kansei-based evaluation process is to identify what kansei features people often use to express their feelings regarding traditional crafts. Each kansei feature is defined by an opposite pair of (adjectival) kansei words,

for example the *fun* feature determines the pair of kansei words *solemn* and *funny*. Let

1.  $\{F_1, \dots, F_K\}$  be the set of kansei features selected,
2.  $\mathbf{w}_k^+$  and  $\mathbf{w}_k^-$  be the opposite pair of kansei words corresponding to  $F_k$ , for  $k = 1, \dots, K$ . Denote  $\mathbf{W}$  the set of kansei words, i.e.  $\mathbf{W} = \{\mathbf{w}_k^+, \mathbf{w}_k^- | k = 1, \dots, K\}$ .

Then, the SD method [8] is used as a measurement instrument to design the questionnaire for gathering kansei evaluation data. Particularly, the questionnaire using the SD method for gathering information consists in listing the kansei features, each of which corresponds to an opposite pair of kansei words that lie at either end of a qualitative  $M$ -point scale, where  $M$  is an odd positive integer as used, for example, in 5-point scale, 7-point scale or 9-point scale. Let us symbolically denote the  $M$ -point scale by

$$\mathbb{V} = \{v_1, \dots, v_M\} \tag{1}$$

where  $\mathbf{w}_k^+$  and  $\mathbf{w}_k^-$  are respectively assumed to be at the ends  $v_1$  and  $v_M$ .

The questionnaire is then distributed to a population  $\mathcal{P}$  of subjects who are invited to express their emotional assessments according each kansei feature of craft patterns in  $\mathcal{O}$  by using the  $M$ -point scale. Formally, we can model the kansei data of each craft pattern  $o_i \in \mathcal{O}$  according to kansei features obtained from the assessment of subjects  $s_j$  in  $\mathcal{P}$  as shown in Table 1, where  $x_{jk}(o_i) \in \mathbb{V}$ , for  $j = 1, \dots, P = |\mathcal{P}|$  and  $k = 1, \dots, K$ .

**Table 1.** The kansei assessment data of pattern  $o_i$

Subjects	Kansei Features			
	$F_1$	$F_2$	$\dots$	$F_K$
$s_1$	$x_{11}(o_i)$	$x_{12}(o_i)$	$\dots$	$x_{1K}(o_i)$
$s_2$	$x_{21}(o_i)$	$x_{22}(o_i)$	$\dots$	$x_{2K}(o_i)$
$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$
$s_P$	$x_{P1}(o_i)$	$x_{P2}(o_i)$	$\dots$	$x_{PK}(o_i)$

The kansei assessment database built, as described above, will be utilized to generate the knowledge serving for the following evaluation problem. Assume that an agent as a potential consumer is interested in looking for a craft pattern which would meet her preference given by a proper subset  $W$  of the set  $\mathbf{W}$  of kansei words as defined below. She may then want to rate craft patterns available in  $\mathcal{O}$  according to her preference. In particular, we are concerned with consumer-specified requests which can be stated generally in forms of the following statement:

“I should like craft items which would best meet  $LQ$  (of) my preference specified in  $W \subset \mathbf{W}$ ” (★)

where  $LQ$  is a linguistic quantifier such as *all*, *most*, *at least half*, *as many as possible*, etc. Formally, the problem can be formulated as follows.

Given  $W = \{\mathbf{w}_{k_1}^*, \dots, \mathbf{w}_{k_n}^*\}$  and  $LQ$  corresponding to the request specified by an agent as linguistically stated in ( $\star$ ), where  $*$  stands for either  $+$  or  $-$ , and  $\{k_1, \dots, k_n\} \subseteq \{1, \dots, K\}$ , the problem now is how to evaluate craft patterns in  $\mathcal{O}$  using kansei data and the request specified as the pair  $[W, LQ]$ ? Here, by  $*$  standing for either  $+$  or  $-$  as above, it indicates that only one of the two  $\mathbf{w}_{k_l}^+$  and  $\mathbf{w}_{k_l}^-$  ( $l = 1, \dots, n$ ) presents in  $W$ , which may be psychologically reasonable to assume. For example, if the agent is interested in craft items being *funny* according to kansei feature of *fun*, then she is not interested in those being *solemn*, the opposite kansei word of *funny*.

In the following section, we will introduce two decision analysis based procedures for solving this consumer-oriented evaluation problem. Namely, the first evaluation procedure is based on fuzzy target-based decision analysis approach, and the second one is based on the linguistic decision analysis approach with the 2-tuple linguistic representation model [12].

### 3 Two Decision Analysis Based Evaluation Procedures

#### 3.1 Evaluation Method Using Target-Based Decision Analysis

Viewing multi-person assessments as uncertain judgments regarding kansei features of traditional craft items, the above-mentioned evaluation problem can be solved by applying the fuzzy target-based decision model recently developed in [5] as follows.

First, let us denote  $\mathbf{D}$  the kansei assessment database about a finite set  $\mathcal{O}$  of craft patterns using SD method as mentioned previously, and  $\mathbf{D}[o_i]$  the data of pattern  $o_i$  ( $i = 1, \dots, N$ ) as shown in Table 1.

For each pattern  $o_i$ , we define for each kansei feature  $F_k$ ,  $k = 1, \dots, K$ , a probability distribution  $f_{ik} : \mathbb{V} \rightarrow [0, 1]$  as follows:

$$f_{ik}(v_h) = \frac{|\{s_j \in \mathcal{P} : x_{jk}(o_i) = v_h\}|}{|\mathcal{P}|} \quad (2)$$

This distribution  $f_{ik}$  is considered as an uncertain judgment of craft pattern  $o_i$  according to kansei feature  $F_k$ . By the same way, we can obtain a  $K$ -tuple of distributions  $[f_{i1}, \dots, f_{iK}]$  regarding the kansei assessment of  $o_i$  and call the tuple the kansei profile of  $o_i$ . Similarly, kansei profiles of all patterns in  $\mathcal{O}$  can be generated from  $\mathbf{D}$ .

Having generated kansei profiles for all patterns  $o_i \in \mathcal{O}$  as above, we now define the evaluation function  $V$  corresponding to the request ( $\star$ ) symbolically specified as  $[W, LQ]$ , where  $W = \{\mathbf{w}_{k_1}^*, \dots, \mathbf{w}_{k_n}^*\}$  and  $LQ$  is a linguistic quantifier.

Intuitively, if a consumer expresses her preference on a kansei feature such as *color contrast* with kansei word *bright*, she may implicitly assume a preference order on the semantic differential scale corresponding to *color contrast* towards

the end  $v_1$  where *bright* is placed. Conversely, if the consumer’s preference on *color contrast* was *dark*, i.e. the opposite kansei word of *bright*, she would assume an inverse preference order on the scale towards the end  $v_M$  where *dark* is placed. In other words, in consumer-oriented evaluation using kansei data, the preference order on the semantic differential scale corresponding to a kansei feature should be determined adaptively depending on a particular consumer’s preference. This can be formally formulated as below.

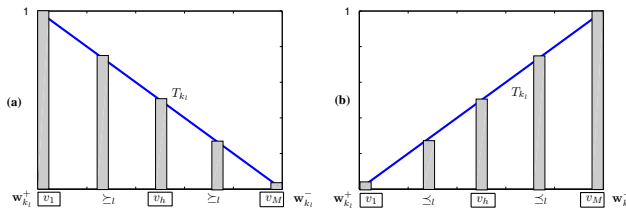
For each  $\mathbf{w}_{k_l}^* \in W$ , we define a linear preference order  $\succeq_l$  on  $\mathbb{V}$  according to the kansei feature  $F_{k_l}$  as follows

$$v_h \succeq_l v_{h'} \Leftrightarrow \begin{cases} h' \geq h, & \text{if } \mathbf{w}_{k_l}^* = \mathbf{w}_{k_l}^+ \\ h \geq h', & \text{if } \mathbf{w}_{k_l}^* = \mathbf{w}_{k_l}^- \end{cases} \quad (3)$$

In addition, due to vagueness inherent in consumer’s expression of preference in terms of kansei words, each  $\mathbf{w}_{k_l}^*$  is considered as the feeling target, denoted by  $T_{k_l}$ , of the consumer according to kansei feature  $F_{k_l}$ , which can be represented as a possibility variable (Zadeh, 1978) on  $\mathbb{V}$  whose possibility distribution is defined as

$$\pi_{k_l}(v_h) = \begin{cases} \left(\frac{M-h}{M-1}\right)^m, & \text{if } \mathbf{w}_{k_l}^* = \mathbf{w}_{k_l}^+ \\ \left(\frac{h-1}{M-1}\right)^m, & \text{if } \mathbf{w}_{k_l}^* = \mathbf{w}_{k_l}^- \end{cases} \quad (4)$$

where  $m \geq 0$  expresses the degree of intensity of the consumer’s feelings about the target. Intuitively, when a consumer expresses her feeling targets using kansei words combined with linguistic modifiers such as *very*, *slightly*, etc., to emphasize her intensity about targets, the degree of intensity  $m$  can then be determined similarly as in Zadeh’s method of modelling linguistic modifiers via power functions in approximate reasoning [16]. Fig. 1 graphically illustrates these concepts for the case  $m = 1$ , which exhibits a neutral-intensity toward targets.



**Fig. 1.** The preference order  $\succeq_l$  and the possibility distribution of feeling target  $T_{k_l}$ : (a)  $\mathbf{w}_{k_l}^* = \mathbf{w}_{k_l}^+$ ; (b)  $\mathbf{w}_{k_l}^* = \mathbf{w}_{k_l}^-$

As such, with the consumer’s preference specified by  $W$ , we obtain  $n$  feeling targets  $T_{k_l}$  ( $l = 1, \dots, n$ ) accompanying with  $n$  preference orders  $\succeq_l$  ( $l = 1, \dots, n$ ) on the semantic differential scale of kansei features  $F_{k_l}$  ( $l = 1, \dots, n$ ), respectively. Recall that, for each  $l = 1, \dots, n$ , the uncertain judgment of each

craft pattern  $o_i$  regarding the kansei feature  $F_{k_l}$  is represented by the probability distribution  $f_{ik_l}$  over  $\mathbb{V}$ , as defined previously. Now we are able to evaluate, for each  $l = 1, \dots, n$ , how the feeling performance of a pattern  $o_i$  on  $F_{k_l}$ , denoted by  $F_{k_l}(o_i)$  and represented by  $f_{ik_l}$ , meets the feeling target  $T_{k_l}$  representing consumer's preference on  $F_{k_l}$ . This can be done as follows.

Firstly, making use of the possibility-probability conversion method [15] we can transform the possibility distribution of feeling target  $T_{k_l}$  into an associated probability distribution, denoted by  $\hat{p}_{k_l}$ , via the simple normalization as follows

$$\hat{p}_{k_l}(v_h) = \frac{\pi_{k_l}(v_h)}{\sum_{v \in \mathbb{V}} \pi_{k_l}(v)} \tag{5}$$

Then, by accepting the assumption that the feeling target  $T_{k_l}$  is stochastically independent of feeling performance on  $F_{k_l}$  of any pattern  $o_i$ , we can work out the probability that the feeling performance  $F_{k_l}(o_i)$  meets the feeling target  $T_{k_l}$ , denoted by  $\mathbf{P}(F_{k_l}(o_i) \succeq T_{k_l})$ , in terms of the preference order  $\succeq_l$  as

$$\begin{aligned} \mathbf{P}(F_{k_l}(o_i) \succeq T_{k_l}) &\triangleq P(f_{ik_l} \succeq_l \hat{p}_{k_l}) \\ &= \sum_{h=1}^M f_{ik_l}(v_h) P(v_h \succeq_l \hat{p}_{k_l}) \end{aligned} \tag{6}$$

where  $P(v_h \succeq_l \hat{p}_{k_l})$  is the cumulative probability function defined by

$$P(v_h \succeq_l \hat{p}_{k_l}) = \sum_{v_h \succeq_l v_{h'}} \hat{p}_{k_l}(v_{h'}) \tag{7}$$

Intuitively, the quantity  $\mathbf{P}(F_{k_l}(o_i) \succeq T_{k_l})$  defined above could be interpreted as the probability of “the feeling performance on  $F_{k_l}$  of  $o_i$  meeting the feeling target  $T_{k_l}$  specified by a consumer on  $F_{k_l}$ ”. Then, after having these probabilities  $\mathbf{P}(F_{k_l}(o_i) \succeq T_{k_l}) = \mathbf{P}_{k_l i}$ , for  $l = 1, \dots, n$ , we are able to aggregate all of them to obtain an aggregated value with taking the linguistic quantifier  $LQ$  into account, making use of the so-called ordered weighted averaging (OWA) aggregation operator [14].

Under such a semantics of OWA operators, now we are ready to define the evaluation function, for any  $o_i \in \mathcal{O}$ , as follows

$$\begin{aligned} V(o_i) &= \mathcal{F}(\mathbf{P}_{k_1 i}, \dots, \mathbf{P}_{k_n i}) \\ &= \sum_{l=1}^n w_l \mathbf{P}_{l i} \end{aligned} \tag{8}$$

where  $\mathbf{P}_{l i}$  is the  $l$ -th largest element in the collection  $\mathbf{P}_{k_1 i}, \dots, \mathbf{P}_{k_n i}$  and weighting vector  $[w_1, \dots, w_n]$  is determined directly by using a fuzzy set-based semantics of the linguistic quantifier  $LQ$ . As interpreted previously on quantities  $\mathbf{P}_{k_l i}$  ( $l = 1, \dots, n$ ), the aggregated value  $V(o_i)$  therefore indicates the degree to which craft pattern  $o_i$  meets the feeling preference derived from the request specified by a consumer as  $[W, LQ]$ .

### 3.2 Evaluation Method Using Linguistic Decision Analysis

Now we will develop in this section another evaluation method, making use of linguistic decision analysis with the 2-tuple linguistic representation model [4]. The main reason for using the 2-tuple based evaluation approach is due to its advantage over conventional fuzzy set-based and symbolic approaches; it overcomes the limitations of the loss of information yielded by the process of linguistic approximation, and the lack of precision in final results inherently faced by these conventional approaches.

To make the 2-tuple linguistic representation model applicable to the evaluation problem at hand, we will treat qualitative assessments regarding each kansei feature given in the 7-point scale as linguistic assessments accordingly taken from the set  $\mathcal{S}$  of seven linguistic terms as described in Fig. 2.

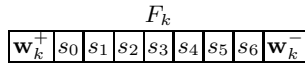


Fig. 2. Linguistic values and their relation to a pair of kansei words

In the 2-tuple representation model, linguistic information is represented by a linguistic 2-tuple  $(s, \alpha)$  composed of a linguistic term  $s \in \mathcal{S}$  and a number  $\alpha \in [-0.5, 0.5]$ . More particularly, let  $\mathcal{S} = \{s_0, \dots, s_g\}$  be a linguistic term set on which a total order is defined as:  $s_i \leq s_j \Leftrightarrow i \leq j$ . In addition, a negation operator  $\text{Neg}$  can be defined by:  $\text{Neg}(s_i) = s_j$  such that  $j = g - i$ . In general, applying a symbolic method for aggregating linguistic information often yields a value  $\beta \in [0, g]$ , and  $\beta \notin \{0, \dots, g\}$ , then a symbolic approximation must be used to get the result expressed in  $\mathcal{S}$ .

In order to avoid any approximation process which causes a loss of information in the processes of computing with words, alternatively the 2-tuple linguistic representation model takes  $\mathcal{S} \times [-0.5, 0.5]$  as the underlying space for representing information. In this representation space, if a value  $\beta \in [0, g]$  represents the result of a linguistic aggregation operation, then the 2-tuple  $(s_i, \alpha)$  that expresses the information equivalent to  $\beta$  is obtained by means of the following transformation:

$$\begin{aligned} \Delta : [0, g] &\longrightarrow \mathcal{S} \times [-0.5, 0.5] \\ \beta &\longmapsto (s_i, \alpha), \end{aligned}$$

with

$$\begin{cases} i = \text{round}(\beta) \\ \alpha = \beta - i \end{cases}$$

and then  $\alpha$  is called a *symbolic translation*, which supports the “difference of information” between the value  $\beta \in [0, g]$  obtained after a symbolic aggregation operation, and the closest value in  $\{0, \dots, g\}$  indicating the index of the best matched term in  $\mathcal{S}$ .



Inversely, a 2-tuple  $(s_i, \alpha) \in \mathcal{S} \times [-0.5, 0.5]$  can also be equivalently represented by a numerical value in  $[0, g]$  by means of the following transformation:

$$\begin{aligned} \Delta^{-1} : \mathcal{S} \times [-0.5, 0.5] &\longrightarrow [0, g] \\ (s_i, \alpha) &\longmapsto \Delta^{-1}(s_i, \alpha) = i + \alpha. \end{aligned}$$

Under such transformations, it should be noticed here that any original linguistic term  $s_i$  in  $\mathcal{S}$  is then represented by its equivalent 2-tuple  $(s_i, 0)$  in the 2-tuple linguistic model.

The comparison of linguistic information represented by 2-tuples is defined as follows. Let  $(s_i, \alpha_1)$  and  $(s_j, \alpha_2)$  be two 2-tuples, then

- if  $i < j$  then  $(s_i, \alpha_1)$  is less than  $(s_j, \alpha_2)$ ,
- if  $i = j$  then
  1. if  $\alpha_1 = \alpha_2$  then  $(s_i, \alpha_1)$  and  $(s_j, \alpha_2)$  represent the same information,
  2. if  $\alpha_1 < \alpha_2$  then  $(s_i, \alpha_1)$  is less than  $(s_j, \alpha_2)$ ,
  3. if  $\alpha_1 > \alpha_2$  then  $(s_i, \alpha_1)$  is greater than  $(s_j, \alpha_2)$ .

Using two 2-tuple transformations defined above, the negation operator over 2-tuples is defined as follows:

$$\text{Neg}((s_i, \alpha)) = \Delta(g - (\Delta^{-1}(s_i, \alpha))) \tag{9}$$

Now the consumer-oriented evaluation method based on the 2-tuple representation model can be formulated as follows.

Given a request  $[W, LQ]$  with  $W = \{\mathbf{w}_{k_1}^*, \dots, \mathbf{w}_{k_n}^*\}$  and  $LQ$  as a linguistic quantifier, let us decompose the set of indices  $I = \{k_1, \dots, k_n\}$  into two disjoint subsets  $I^+$  and  $I^-$  such that

$$I^+ = \{k_j \in I \mid \mathbf{w}_{k_j}^* = \mathbf{w}_{k_j}^+\} \text{ and } I^- = \{k_j \in I \mid \mathbf{w}_{k_j}^* = \mathbf{w}_{k_j}^-\} \tag{10}$$

Then, for each object  $o_i \in \mathcal{O}$ , the performance of  $o_i$  on the kansei feature  $F_{k_j}$  is evaluated by

$$V_{k_j}(o_i) = \Delta \left( \sum_{s \in \mathcal{S}} f_{ik_j}(s) \Delta^{-1}(s, 0) \right), \text{ if } k_j \in I^- \tag{11}$$

and

$$V_{k_j}(o_i) = \Delta \left( \sum_{s \in \mathcal{S}} f_{ik_j}(s) \Delta^{-1}(\text{Neg}((s, 0))) \right), \text{ if } k_j \in I^+ \tag{12}$$

where  $f_{ik_j}(s)$  is defined by

$$f_{ik_j}(s) = \frac{|\{s_h \in \mathcal{P} : x_{hk_j}(o_i) = s\}|}{|\mathcal{P}|} \tag{13}$$

That is,  $V_{k_j}(o_i)$  is the mean value of uncertain linguistic assessment of  $o_i$  regarding the kansei feature  $F_{k_j}$  computed by means of linguistic 2-tuples. Once

values  $V_{k_j}(o_i)$  have been computed for all features  $F_{k_j}, k_j \in I$ , the overall performance of  $o_i$  is calculated by aggregating all of them using an OWA operator  $\mathcal{F}$  of dimension  $n$  similar to (8), such as

$$V(o_i) = \mathcal{F}(V_{k_1}(o_i), \dots, V_{k_n}(o_i)) \quad (14)$$

with the associated weighting vector  $[w_1, \dots, w_n]$  determined by using the fuzzy set-based semantics of linguistic quantifier  $LQ$ .

## 4 Illustration to Evaluation of Kutani Coffee Cups

For illustration, we shall apply the proposed model to evaluating Kutani porcelain, a traditional craft industry in Japan, historically back to the seventeenth century, of Kutani Pottery Village in Ishikawa prefecture<sup>1</sup>. A total of 35 patterns of Kutani coffee cup have been collected for the Kansei-based evaluation, and 26 opposite pairs of kansei words were used to design the answer sheet for gathering kansei assessment data of these items (i.e., Kutani coffee cups) for evaluation. Kansei words are approximately translated into English as shown in Table 2.

A total of 60 subjects were invited to participate in the kansei assessment process. The data obtained is 3-way data of which each Kutani coffee cup  $\#i$  ( $i = 1, \dots, 35$ ) is assessed by all participated subjects on all kansei features  $F_k, k = 1, \dots, 26$ . The 3-way data is then used to generate kansei profiles for patterns via (2) as mentioned previously. These kansei profiles are considered as (uncertain) feeling assessments of items serving as the knowledge for consumer-oriented evaluation.

### 4.1 Obtained Results of Two Methods

Assuming a consumer's request is specified as

$$\{[\mathbf{w}_7^+, \mathbf{w}_{10}^-, \mathbf{w}_{11}^+, \mathbf{w}_{17}^+, \mathbf{w}_{25}^-], \text{ as many as possible}\}$$

That is, verbally, she would ask for items meeting *as many as possible* of her feeling preferences of *pretty, attractive, flowery, bright* and *pale*.

We first determine preference orders on  $\mathbb{V} = \{v_1, \dots, v_7\}$  for features  $F_7, F_{10}, F_{11}, F_{17}$  and  $F_{25}$ . Using (3), we have  $\succeq_{10} = \succeq_{25}$  and  $\succeq_7 = \succeq_{11} = \succeq_{17}$ , where

$$v_1 \succeq_7 \dots \succeq_7 v_7 \text{ and } v_7 \succeq_{10} \dots \succeq_{10} v_1$$

Then, using (4) for  $m = 2$ , we define feeling targets  $T_7, T_{10}, T_{11}, T_{17}$  and  $T_{25}$  for features  $F_7, F_{10}, F_{11}, F_{17}$  and  $F_{25}$  respectively. In this case we have  $T_{10} \equiv T_{25}$  and  $T_7 \equiv T_{11} \equiv T_{17}$  with possibility distributions shown in Fig. 3.

We now determine the weighting vector of dimension 5, denoted by  $w = [w_1, w_2, w_3, w_4, w_5]$ , according to the fuzzy set-based semantics of linguistic quantifier '*as many as possible*'. Assume that, for example, the membership function

<sup>1</sup> [http://shofu.pref.ishikawa.jp/shofu/intro\\_e/HTML/H\\_S50402.html](http://shofu.pref.ishikawa.jp/shofu/intro_e/HTML/H_S50402.html)

**Table 2.** Opposite pairs of kansei words used for the evaluation

$F_k$	Left kansei word	$v_1$	$\dots$	$v_7$	Right kansei word
1	conventional( $\mathbf{w}_1^+$ )	<input type="checkbox"/>	$\dots$	<input type="checkbox"/>	unconventional( $\mathbf{w}_1^-$ )
2	simple( $\mathbf{w}_2^+$ )	<input type="checkbox"/>	$\dots$	<input type="checkbox"/>	compound( $\mathbf{w}_2^-$ )
3	solemn( $\mathbf{w}_3^+$ )	<input type="checkbox"/>	$\dots$	<input type="checkbox"/>	funny( $\mathbf{w}_3^-$ )
4	formal( $\mathbf{w}_4^+$ )	<input type="checkbox"/>	$\dots$	<input type="checkbox"/>	causal( $\mathbf{w}_4^-$ )
5	serene( $\mathbf{w}_5^+$ )	<input type="checkbox"/>	$\dots$	<input type="checkbox"/>	forceful( $\mathbf{w}_5^-$ )
6	still( $\mathbf{w}_6^+$ )	<input type="checkbox"/>	$\dots$	<input type="checkbox"/>	moving( $\mathbf{w}_6^-$ )
7	pretty( $\mathbf{w}_7^+$ )	<input type="checkbox"/>	$\dots$	<input type="checkbox"/>	austere( $\mathbf{w}_7^-$ )
8	friendly( $\mathbf{w}_8^+$ )	<input type="checkbox"/>	$\dots$	<input type="checkbox"/>	unfriendly( $\mathbf{w}_8^-$ )
9	soft( $\mathbf{w}_9^+$ )	<input type="checkbox"/>	$\dots$	<input type="checkbox"/>	hard( $\mathbf{w}_9^-$ )
10	blase( $\mathbf{w}_{10}^+$ )	<input type="checkbox"/>	$\dots$	<input type="checkbox"/>	attractive( $\mathbf{w}_{10}^-$ )
11	flowery( $\mathbf{w}_{11}^+$ )	<input type="checkbox"/>	$\dots$	<input type="checkbox"/>	quiet( $\mathbf{w}_{11}^-$ )
12	happy( $\mathbf{w}_{12}^+$ )	<input type="checkbox"/>	$\dots$	<input type="checkbox"/>	normal( $\mathbf{w}_{12}^-$ )
13	elegant( $\mathbf{w}_{13}^+$ )	<input type="checkbox"/>	$\dots$	<input type="checkbox"/>	loose( $\mathbf{w}_{13}^-$ )
14	delicate( $\mathbf{w}_{14}^+$ )	<input type="checkbox"/>	$\dots$	<input type="checkbox"/>	large-hearted( $\mathbf{w}_{14}^-$ )
15	luxurious( $\mathbf{w}_{15}^+$ )	<input type="checkbox"/>	$\dots$	<input type="checkbox"/>	frugal( $\mathbf{w}_{15}^-$ )
16	gentle( $\mathbf{w}_{16}^+$ )	<input type="checkbox"/>	$\dots$	<input type="checkbox"/>	pithy( $\mathbf{w}_{16}^-$ )
17	bright( $\mathbf{w}_{17}^+$ )	<input type="checkbox"/>	$\dots$	<input type="checkbox"/>	dark( $\mathbf{w}_{17}^-$ )
18	reserved( $\mathbf{w}_{18}^+$ )	<input type="checkbox"/>	$\dots$	<input type="checkbox"/>	imperious( $\mathbf{w}_{18}^-$ )
19	free( $\mathbf{w}_{19}^+$ )	<input type="checkbox"/>	$\dots$	<input type="checkbox"/>	regular( $\mathbf{w}_{19}^-$ )
20	level( $\mathbf{w}_{20}^+$ )	<input type="checkbox"/>	$\dots$	<input type="checkbox"/>	indented( $\mathbf{w}_{20}^-$ )
21	lustered( $\mathbf{w}_{21}^+$ )	<input type="checkbox"/>	$\dots$	<input type="checkbox"/>	matte( $\mathbf{w}_{21}^-$ )
22	transpicious( $\mathbf{w}_{22}^+$ )	<input type="checkbox"/>	$\dots$	<input type="checkbox"/>	dim( $\mathbf{w}_{22}^-$ )
23	warm( $\mathbf{w}_{23}^+$ )	<input type="checkbox"/>	$\dots$	<input type="checkbox"/>	cool( $\mathbf{w}_{23}^-$ )
24	moist( $\mathbf{w}_{24}^+$ )	<input type="checkbox"/>	$\dots$	<input type="checkbox"/>	arid( $\mathbf{w}_{24}^-$ )
25	colorful( $\mathbf{w}_{25}^+$ )	<input type="checkbox"/>	$\dots$	<input type="checkbox"/>	sober( $\mathbf{w}_{25}^-$ )
26	plain( $\mathbf{w}_{26}^+$ )	<input type="checkbox"/>	$\dots$	<input type="checkbox"/>	gaudy, loud( $\mathbf{w}_{26}^-$ )

of the quantifier ‘as many as possible’ is defined as a mapping  $Q : [0, 1] \rightarrow [0, 1]$  such that

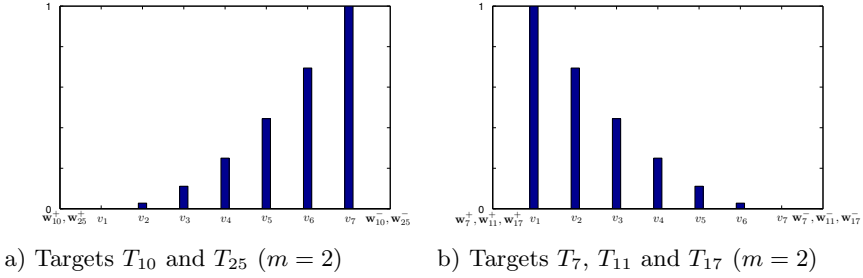
$$Q(r) = \begin{cases} 0 & \text{if } 0 \leq r \leq 0.5 \\ 2r - 1 & \text{if } 0.5 \leq r \leq 1 \end{cases}$$

We then obtain the weighting vector as  $w = [0, 0, 0.2, 0.4, 0.4]$  using Yager’s method proposed in [14].

With these preparations we are now ready to apply the target based evaluation method for ranking items Kutani\_cup# $i$ ,  $i = 1, \dots, 35$ , as follows. First, we use (5) and (6) for computing probabilities  $\mathbf{P}_{7i}$ ,  $\mathbf{P}_{10i}$ ,  $\mathbf{P}_{11i}$ ,  $\mathbf{P}_{17i}$  and  $\mathbf{P}_{25i}$  of meeting corresponding feeling targets  $T_7$ ,  $T_{10}$ ,  $T_{11}$ ,  $T_{17}$  and  $T_{25}$  for each item Kutani\_cup# $i$  ( $i = 1, \dots, 35$ ). Then, using (8) we have

$$V(\text{Kutani\_cup}\#i) = \mathcal{F}(\mathbf{P}_{7i}, \mathbf{P}_{10i}, \mathbf{P}_{11i}, \mathbf{P}_{17i}, \mathbf{P}_{25i})$$

where  $\mathcal{F}$  is the OWA operator of dimension 5 associated with the weighting vector  $w = [0, 0, 0.2, 0.4, 0.4]$ .



**Fig. 3.** Possibility distribution of feeling targets

Table 3 shows the top three patterns that would best meet the feeling preferences *pretty*, *attractive*, *flowery*, *bright* and *pale*, with different typical linguistic quantifiers used.

**Table 3.** Quantifiers used and corresponding top 3 patterns

Linguistic quantifier	Weighting vector	The top 3 patterns
<i>As many as possible</i> (AMAP)	[0, 0, 0.2, 0.4, 0.4]	#4 $\succeq$ #8 $\succeq$ #11
<i>All</i>	[0, 0, 0, 0, 1]	#8 $\succeq$ #7 $\succeq$ #30
<i>There exists</i>	[1, 0, 0, 0, 0]	#13 $\succeq$ #18 $\succeq$ #29
<i>At least haft</i> (ALH)	[0.4, 0.4, 0.2, 0, 0]	#13 $\succeq$ #6 $\succeq$ #24

On the other hand, using the 2-tuple based evaluation method described above, we also obtain results of the top 3 recommended items with different linguistic quantifiers applied as shown in Table 4.

**Table 4.** Top 3 items recommended using the 2-tuple based method

Linguistic quantifier	Weighting vector	The top 3 patterns
<i>As many as possible</i> (AMAP)	[0, 0, 0.2, 0.4, 0.4]	#8 $\succeq$ #11 $\succeq$ #4
<i>All</i>	[0, 0, 0, 0, 1]	#7 $\succeq$ #9 $\succeq$ #8
<i>There exists</i>	[1, 0, 0, 0, 0]	#13 $\succeq$ #29 $\succeq$ #18
<i>At least haft</i> (ALH)	[0.4, 0.4, 0.2, 0, 0]	#13 $\succeq$ #6 $\succeq$ #24

## 4.2 Comparative Analysis

For the sake of facilitating the discussion of obtained results, all the recommended items by the target based evaluation method (according to typical linguistic quantifiers used) as well as their uncertain assessments on selected features  $F_7, F_{10}, F_{11}, F_{17}$  and  $F_{25}$  are depicted in Fig. 4.

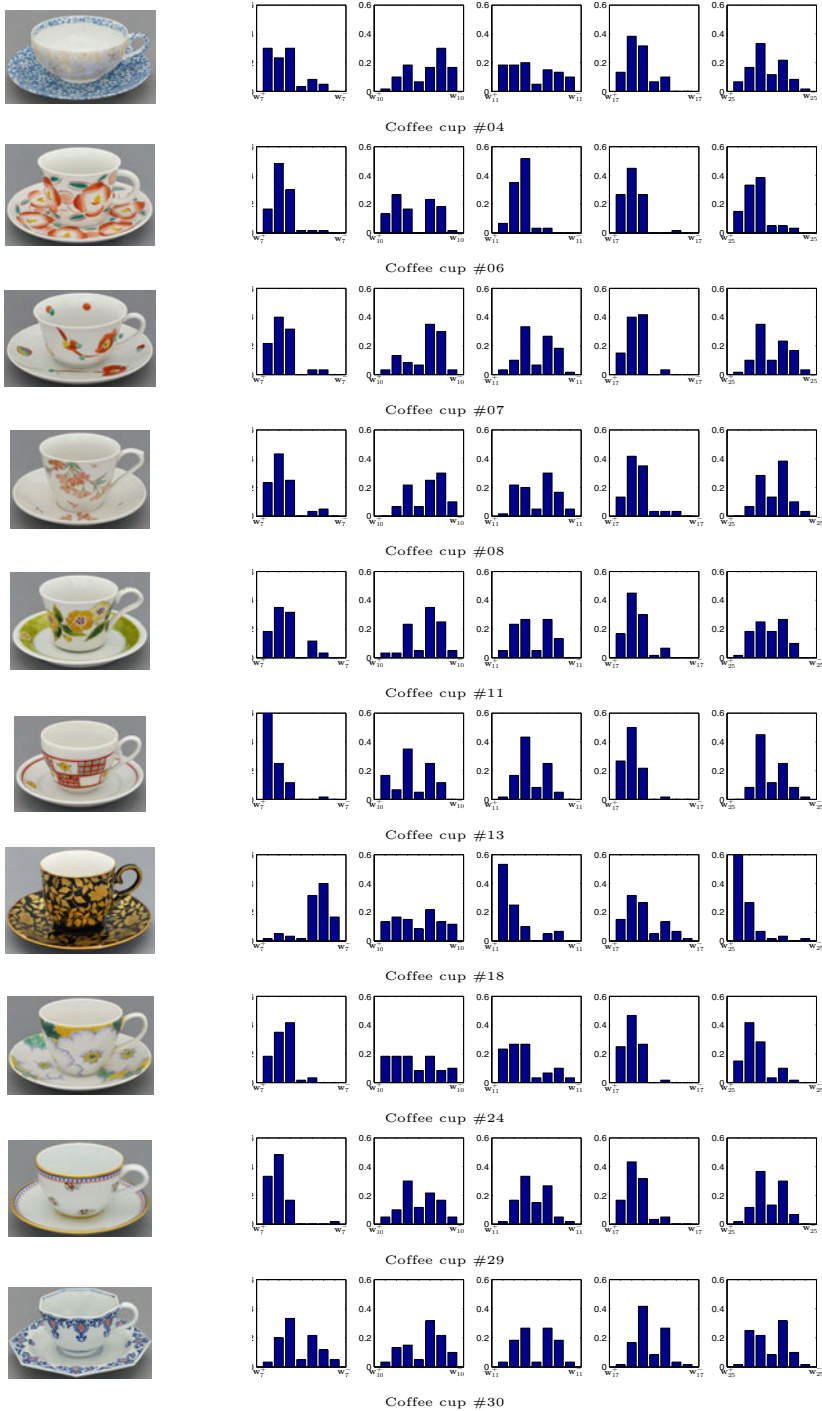


Fig. 4. Recommended items and those uncertain judgments for selected features

As we have seen from Tables 3 and 4, the results yielded by two methods are quite different, except the case of quantifier ‘at least half’. Particularly, in the first case of quantifier ‘as many as possible’, though two methods produced the same top 3 items but these items were ranked differently. Item #4 was ranked first by the target-based method while it appears as the third by the 2-tuple based method. For the case of quantifier ‘all’, it is worth noting here that the uncertain judgments of items #7 and #8 on correspondingly selected features are somewhat similar. However, item #8 was ranked third by the 2-tuple based method and dominated by item #9 as the second and item #7 as the first, while it was ranked first by the target-based method. In the case of quantifier ‘there exists’, a position interchange of items #18 and #29 happens, in particular item #18 was dominated by item #29 with the 2-tuple based method and vice-versa with the target-based method. In fact, the target achievement of item #18 on target *flowery* ( $w_{11}^+$ ) is 0.7209 which is better than that of item #29 on target *pretty* ( $w_7^+$ ) as 0.6804. This can be observed as illustrated in Fig. 4.

The difference in ranking results of the two methods can be explained as follows. In the 2-tuple based method, only preferences over the linguistic term set  $\mathcal{S}$  induced from the consumer’s request are taken into account. While in the target-based method, not only these preferences but also feeling targets specified by the consumer are considered simultaneously. From a decision analysis point of view, after determining consumer-specified preferences the 2-tuple based method applies the expected value model (refer to (11) and (12)) to evaluate the performance of an object regarding each kansei feature specified by the consumer. Thus, as discussed in Huynh *et al.* [5], the 2-tuple based method works similarly to the target-based method when the ‘neutral target’ is used. In particular, if we define targets as

$$\pi_{k_l}(v_h) = 1$$

instead of the targets defined in (4), then the result obtained by the target-based method is the same as that produced by the 2-tuple based method. This means that the target-based method can provide recommendations which would interestingly reflect attitudes of consumer towards feeling targets as graphically illustrated by Fig. 4, whilst those recommended by the 2-tuple based method would not do so.

## 5 Conclusions

In this paper we have conducted a comparative study of two decision analysis based evaluation methods for the evaluation problem of Japanese traditional crafts, which take consumer-specified preferences on kansei features of traditional products into consideration. In doing so, we have first customized the linguistic 2-tuple representation model in linguistic decision analysis in order to apply it to the consumer-oriented evaluation problem using kansei data. It has been shown that the 2-tuple based evaluation method can be seen as a special case of the target-based evaluation method which would interestingly provide the evaluated results reflecting attitudes of consumers about feeling targets.

## References

1. Adomavicius, G., Tuzhilin, A.: Toward the next generation of recommender systems: A survey of the state-of-the-art and possible extensions. *IEEE Trans. Know. Data Eng.* 17, 734–749 (2005)
2. Bordley, R., Kirkwood, C.: Multiattribute preference analysis with performance targets. *Operations Research* 52, 823–835 (2004)
3. Manouselis, N., Costopoulou, C.: Analysis and classification of multi-criteria recommender systems. *World Wide Web* 10, 415–441 (2007)
4. Herrera, F., Martínez, L.: A 2-tuple fuzzy linguistic representation model for computing with words. *IEEE Trans. Fuzzy Syst.* 8, 746–752 (2000)
5. Huynh, V.N., Nakamori, Y., Ryoike, M., Ho, T.B.: Decision making under uncertainty with fuzzy targets. *Fuzzy Optim. Decis. Making* 6, 255–278 (2007)
6. Huynh, V.N., Yan, H.B., Nakamori, Y.: A target-based decision making approach to consumer-oriented evaluation model for Japanese traditional crafts. *IEEE Trans. Eng. Manag.* (in press)
7. Nagamachi, M.: Kansei Engineering: a new ergonomic consumer-oriented technology for product development. *Inter. J. Indus. Ergo.* 15, 3–11 (1995)
8. Osgood, C.E., Suci, G.J., Tannenbaum, P.H.: *The Measurement of Meaning*. University of Illinois Press, Urbana (1957)
9. Petiot, J.-F., Yannou, B.: Measuring consumer perceptions for a better comprehension, specification and assessment of product semantics. *Inter. J. Indus. Ergo.* 33, 507–525 (2004)
10. Ruan, D., Zeng, X. (eds.): *Intelligent Sensory Evaluation: Methodologies and Applications*. Springer, Berlin (2004)
11. Schütte, S.: *Engineering Emotional Values in Product Design – Kansei Engineering in Development*. Linköpings Universitet, Dissertation 951 (2005)
12. Martínez, L.: Sensory evaluation based on linguistic decision analysis. *IJAR* 44, 148–164 (2007)
13. Yan, H.B., Huynh, V.N., Nakamori, Y.: A probability-based approach to consumer oriented evaluation of traditional craft items using kansei data. In: Huynh, V.N., et al. (eds.) *Interval/Probabilistic Uncertainty and Non-Classical Logics*, pp. 326–340. Springer, Heidelberg (2008)
14. Yager, R.R.: On ordered weighted averaging aggregation operators in multicriteria decision making. *IEEE Trans. Syst., Man, Cybern.* 18, 183–190 (1988)
15. Yager, R.R.: On the instantiation of possibility distributions. *Fuzzy Sets Syst.* 128, 261–266 (2002)
16. Zadeh, L.A.: The concept of a linguistic variable and its applications to approximate reasoning. *Inf. Sci.* 8, 199–249, 310–357 (1975)

# Optimization of Multiple Related Negotiation through Multi-Negotiation Network

Fenghui Ren<sup>1,\*</sup>, Minjie Zhang<sup>1</sup>, Chunyan Miao<sup>2</sup>, and Zhiqi Shen<sup>3</sup>

<sup>1</sup> School of Computer Science and Software Engineering  
University of Wollongong, Australia  
{fr510,minjie}@uow.edu.au

<sup>2</sup> School of Computer Engineering

<sup>3</sup> School of Electrical and Electronic Engineering  
Nanyang Technological University, Singapore  
{ASCYMiao,zqshen}@ntu.edu.sg

**Abstract.** In this paper, a Multi-Negotiation Network (MNN) and a Multi-Negotiation Influence Diagram (MNID) are proposed to optimally handle Multiple Related Negotiations (MRN) in a multi-agent system. Most popular, state-of-the-art approaches perform MRN sequentially. However, a sequential procedure may not optimally execute MRN in terms of maximizing the global outcome, and may even lead to unnecessary losses in some situations. The motivation of this research is to use a MNN to handle MRN concurrently so as to maximize the expected utility of MRN. Firstly, both the joint success rate and the joint utility by considering all related negotiations are dynamically calculated based on a MNN. Secondly, by employing a MNID, an agent's possible decision on each related negotiation is reflected by the value of expected utility. Lastly, through comparing expected utilities between all possible policies to conduct MRN, an optimal policy is generated to optimize the global outcome of MRN. The experimental results indicate that the proposed approach can improve the global outcome of MRN in a successful end scenario, and avoid unnecessary losses in an unsuccessful end scenario.

## 1 Introduction

Negotiation is a significant methodology for autonomous agents to reach mutual beneficial agreements in multi-agent systems [1]. People can study and classify negotiations through different aspects. For instance, by considering the number of negotiated issues, negotiations can be classified as *single issue negotiations* and *multiple issue negotiations* [1]. In general, *single issue negotiations* focus on bargains involving only one attribute, while *multiple issue negotiations* contain more than one negotiated issues. By considering the number of negotiators, negotiations can also be classified as *bilateral negotiations* and *multilateral negotiations* [2]. A *bilateral negotiation* is performed between only two negotiators, while a *multilateral negotiation* considers opportunity and competition from other negotiators. By considering negotiation environments, negotiations can be classified as *static negotiations* and *dynamic negotiations* [3]. In a *static negotiation*, the negotiation environment is relatively fixed, and can be fully observed and

---

\* The primary author is a Ph.D candidate.



expected by negotiators. While in a *dynamic negotiation*, the negotiation environment is changed out of a negotiator’s control and can only be partially observed and expected by negotiators. Also, some hybrid partitions combine the above criteria together. For example, Sycara et al. [2] introduced a three-level nest view on negotiations.

Through our studies, it is found that all of these classifications focus only on a negotiation with a sole goal, but do not consider Multiple Related Negotiations (MRN) with different goals. However, in complex negotiation environments, one agent may perform more than one negotiation with different opponents for different goals at same time. Sometimes, these goals are not independent, and these MRN are somehow related. For instance, in a scheduling problem, the negotiation result on the deadline of an early occurring event will definitely impact the negotiation on the starting time of a later occurring event. The negotiation result between a mortgagor and a banker on a mortgage will determine the mortgagor’s reservation in the negotiation with a real estate agent on a property’s price. In order to cover related negotiations, we introduce a three-level hierarchic model in Figure 1 to represent different types of negotiations.

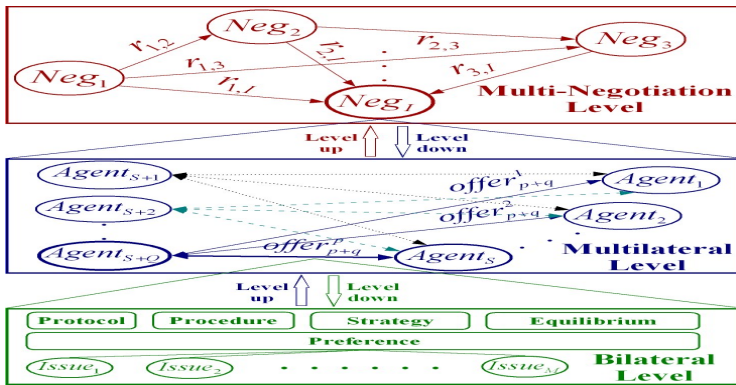


Fig. 1. Three levels hierarchic view of agent negotiation

The first level is named as the *Bilateral Level*, which covers static, bilateral, and multiple issues negotiations. In this level, agents focus on sophisticated negotiation with only one opponent. In order to achieve an optimal outcome, agents may adopt different procedures [1], strategies [4], equilibriums [5] and preferences [6] in negotiation based on individual interest. The second level is named as the *Multilateral Level*, which covers dynamic, multilateral, and multiple issue negotiations. In this level, agents negotiate with more than one opponent synchronously. Researches on this level may pay attention to negotiation pattern selections [7], multilateral negotiation protocols [8] and negotiations in open and dynamic environments [3]. Both the first and second levels focus on one negotiation with a solo goal. The third level is named as the *Multi-Negotiation Level*, which normally pays attention to MRN. Negotiations in this level have different goals, and the goals are somehow related. Each single negotiation in the third level can be represented by a *Bilateral Level* negotiation or a *Multilateral Level* negotiation.

Significant achievements have been reached in agent negotiation in the first two levels with state-of-the-art techniques and approaches. However, very few of them

consider the third level and can handle MRN properly [9,10]. Most existing approaches just separate these MRN and treat each of them individually. The major disadvantage of dealing with those negotiations separately is that if related negotiations are not considered together, the negotiation outcomes may not be optimized or even be damaged for some cases. In order to solve the problem mentioned in MRN, we introduce a Multi-Negotiation Network (MNN) and a Multi-Negotiation Influence Diagram (MNID) in this paper. Firstly, MRN are represented by MNN. Secondly, MNN is extended to MNID. The joint success rate and the joint utility by considering all related negotiations in the MNID is calculated. Thirdly, an optimal policy to conduct MRN is calculated for the MNID, by considering both the joint success rate and the joint utility, to optimize the negotiation outcome of MNID.

## 2 A Multi-Negotiation Network

### 2.1 Construction of a MNN

In this subsection, we introduce notations and the procedure to construct a MNN based on Bayesian Networks [11]. Let a four-tuple  $\langle \mathbb{G}, \mathbf{R}, \mathbf{P}, \Phi \rangle$  indicate a MNN, where  $\mathbb{G} = (\mathbf{V}, \mathbf{E})$  is a directed acyclic graph, set  $\mathbf{R}$  indicates the restriction function between two related negotiations,  $\mathbf{P}$  is a set of success rates, and  $\Phi$  is a set of the utility functions.  $\mathbf{V}$  is a finite, nonempty set of vertices and each vertex indicates a negotiation in a MNN, and  $\mathbf{E}$  is a set of ordered pairs of distinct elements of  $\mathbf{V}$ . Each element of  $\mathbf{E}$  is called a restriction edge with a direction to represent a dependency relationship of two related negotiations, (i.e. a link with arrow between two vertices in a MNN). For example, if a pair  $(V_i, V_j) \in \mathbf{E}$ , we say that there is an edge from  $V_i$  to  $V_j$ . In other words,  $V_j$  depends on  $V_i$ , and  $V_i$  is one of  $V_j$ 's parent. We use function  $r_{ij} : \Phi_j \rightarrow \Phi_j$  ( $r_{ij} \in \mathbf{R}$ ,  $\Phi_j \in \Phi$ ) to indicate the restriction between  $V_i$  and  $V_j$ . If there is no restriction between  $V_i$  and  $V_j$ ,  $r_{ij}$  is *nil*. (That means two negotiations  $V_i$  and  $V_j$  are independent and there is no impact on each other.)  $p_i = P(V_i|pa(V_i))$  ( $p_i \in \mathbf{P}$ ,  $p_i \in [0, 1]$ ) indicates the success rate of  $V_i$ , where  $pa(V_i)$  are the parents of  $V_i$ .  $\Phi_i$  ( $\Phi_i \in \Phi$ ) indicates the utility function of Negotiation  $V_i$ . Figure 2 is an example of a MNN. In this example, the set of vertices in the MNN is  $\mathbf{V} = \{X, Y, Z, W\}$ , the set of edges is  $\mathbf{E} = \{(X, Y), (X, Z), (Y, Z), (Y, W), (Z, W)\}$ , the set of restriction functions is  $\mathbf{R} = \{r_{xy}, r_{xz}, r_{yz}, r_{yw}, r_{zw}\}$ , the set of probabilities is  $\mathbf{P} = \{P(X), P(Y|X), P(Z|X, Y), P(W|Y, Z)\}$ , and the set of utility functions is  $\Phi = \{\Phi_x, \Phi_y, \Phi_z, \Phi_w\}$ .

From a MNN, an agent can view its related negotiations based on dependency relationships of these negotiations. The basic procedure to construct a MNN includes the following three steps.

**Step 1:** to represent each negotiation by a unique vertex  $V_i$  ( $V_i \in \mathbf{V}$ ) and to assign a utility function  $\Phi_i$  ( $\Phi_i \in \Phi$ ) for  $V_i$ . Of course, the agent can modify the utility function anytime;

**Step 2:** to generate all restriction edges, which belong to  $\mathbf{E}$ , between each two related negotiations; and to define restriction functions  $\mathbf{R}$  in the form of  $r_{ij} : \Phi_j \rightarrow \Phi_j$  ( $r_{ij} \in \mathbf{R}$ ).

The restriction function indicates how an ongoing or accomplished negotiation impacts another ongoing negotiation. An accomplished negotiation will not be impacted by other negotiations anymore. For instance, if a buyer synchronously performs two negotiations between a banker and a real estate agent under the condition that the buyer’s reservation on a property’s price depends on the mortgage, then there is a restriction edge from the *mortgage negotiation* to the *property negotiation*. If the negotiation between the buyer and the banker is completed first, its impact on the negotiation between the buyer and the real estate agent will be fixed. However, if the negotiation between the buyer and the real estate agent is completed firstly, the *mortgage negotiation* will not have any further impact on the *property negotiation*. In a MNN, if Negotiation  $V_i$  is an independent negotiation, other negotiations will have no impact on its utility function  $\Phi_i$ . Otherwise, if Negotiation  $V_i$  has  $K$  dependent negotiations, then its utility function will be modified by the consideration of all impacts from these  $K$  dependent negotiations as follows:

$$r_{1i} \circ \dots \circ r_{Ki} : \Phi_i \rightarrow \Phi_i \tag{1}$$

Let us take the MNN in Figure 2 as an example. Negotiation  $X$  has no dependent negotiations, its utility function does not need a modification; Negotiation  $Y$  is dependent on Negotiation  $X$ , its utility function is modified as  $\Phi_y = r_{xy}(\Phi_x)$ ; Negotiation  $Z$  is dependent on Negotiations  $X$  and  $Y$ , its utility function is modified as  $\Phi_z = r_{xz}(r_{yz}(\Phi_z))$ ; and Negotiation  $W$  is dependent on Negotiation  $Y$  and  $Z$ , its utility function is modified as  $\Phi_w = r_{yw}(r_{zw}(\Phi_w))$ .

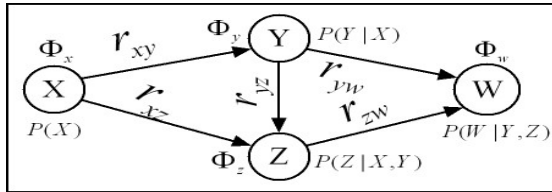


Fig. 2. A Multi-Negotiation Network

**Step 3:** to define the success rate  $p_i$  ( $p_i \in \mathbf{P}$ ) for Negotiation  $V_i$ .

The success rate  $p_i$  indicates how likely an agent’s latest offer will be accepted by its opponents in the remaining negotiation rounds. Suppose that in negotiation round  $t$ , an agent’s latest offer is represented as a utility vector  $(\Phi_i(t), \Phi_i(t)^o)$ , and one of its opponents’ offers is a utility vector  $(u_o^t, u_a^t)$ . The agent’s latest offer generates a payoff of  $\Phi_i(t)$  for itself and  $\Phi_i(t)^o$  for the opponents; and the opponent’s offer generates a payoff of  $u_o^t$  for itself and  $u_a^t$  for the agent. Let  $u_w$  denote the worst possible utility, (a *conflict utility*) for the agent. If the subjective probability of the agent obtaining  $u_w$  is  $p_w$ , then the agent will insist on its offer when its expected utility is greater than the opponent’s offer, ie.,

$$[(1 - p_w)\Phi_i(t) + p_w u_w] \geq u_a^t \tag{2}$$

According to the above inequality, the highest conflict probability that the agent may encounter with the opponent in the next negotiation round is the maximum value of  $p_w$  as follows:

$$p_w = \frac{\Phi_i(t) - u_a^t}{\Phi_i(t) - u_w} \tag{3}$$

Let  $\tau$  be the negotiation deadline and  $t$  be the current round, then the conflict probability that the agent may encounter with the opponent by considering all remaining rounds can be estimated as follows:

$$p_w = \left( \frac{\Phi_i(t) - u_a^t}{\Phi_i(t) - u_w} \right)^{\tau-t} \tag{4}$$

Consequently, the aggregated conflict probability that the agent may encounter before the deadline by considering all opponents in Negotiation  $V_i$  is:

$$p_a = \left( \frac{\prod_{s=1}^{S_i} (\Phi_i(t) - u_s^t)}{(\Phi_i(t) - u_w)^{S_i}} \right)^{\tau-t} \tag{5}$$

where  $S_i$  is the number of opponents in Negotiation  $V_i$ . Therefore, for Negotiation  $V_i$ , the worst success rate  $p_i$  that the agent's offer ( $\Phi_i(t)$ ) will be accepted by at least one opponent before the deadline is:

$$p_i = 1 - p_a = 1 - \left( \frac{\prod_{s=1}^{S_i} (\Phi_i(t) - u_s^t)}{(\Phi_i(t) - u_w)^{S_i}} \right)^{\tau-t} \tag{6}$$

In this subsection, we introduced the concept and notations for a MNN and steps to construct a MNN. It must be pointed out that a MNN can be dynamically modified according to changes of the negotiation environment. In the following subsections, we explain how to dynamically update a MNN.

### 2.2 Updating of a MNN

Since negotiation environments can be highly complex and dynamic in real-world situations, agents may need some modifications on their MRN in order to respond to changes in negotiation environments. Such modifications may include the following cases: starting a new negotiation, terminating an ongoing negotiation, adjusting utility functions, adjusting restriction functions, changing negotiation opponents etc. When

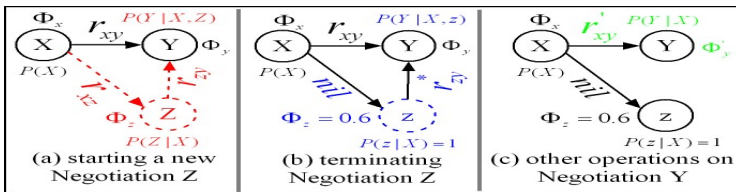


Fig. 3. Multi-Negotiation Network update

these changes happen, agents should immediately update their MNNs. In this subsection, we introduce two major operations and suggest other operations incorporating several major changes on MNN updating.

**Starting a Negotiation.** Assume that there are  $i$  related negotiations. If a new negotiation is commenced by an agent, a vertex  $V_{i+1}$  should be inserted into the MNN to indicate the new negotiation. Also, the agent should define a utility function  $\Phi_{i+1}$  for Negotiation  $V_{i+1}$ , and specify restriction edges between all existing negotiations and Negotiation  $V_{i+1}$ . If there is a restriction edge from an existing Negotiation  $V_i$  to the new Negotiation  $V_{i+1}$ , restriction functions  $r_{i(i+1)}$  should be specified, and the utility function  $\Phi_{i+1}$  should be modified according to this restriction. If there is a restriction edge from the new Negotiation  $V_{i+1}$  to an existing Negotiation  $V_i$ , then Negotiation  $V_i$ 's success rate  $p_i$  and utility function  $\Phi_i$  should also be updated. In Figure 3(a), an example of adding new Negotiation  $Z$  in a MNN is demonstrated.

**Terminating a Negotiation.** If an ongoing Negotiation  $V_i$  is terminated by an agent, no matter whether Negotiation  $V_i$  is successful or failed to reach an agreement, we use lower case letters on Negotiation  $V_i$ 's caption to indicate that the negotiation is in a final state. Meanwhile, the success rate for Negotiation  $V_i$  is set to 1 for a successful negotiation or to 0 for a failed negotiation. For any Negotiation  $V_j$  which Negotiation  $V_i$  depends on, the restriction function  $r_{ji}$  is set to *nil* and Negotiation  $V_i$ 's utility function  $\Phi_i$  is replaced by a constant to indicate the payoff of Negotiation  $V_i$ . For any Negotiation  $V_k$  which depends on Negotiation  $V_i$ , the restriction function  $r_{ik}$  is eventually fixed and its impact on Negotiation  $V_k$ 's utility function is also fixed. In Figure 3(b), an example of terminating an ongoing Negotiation  $Z$  in a MNN is demonstrated.

**Other Operations.** Besides the previous two situations, agents may modify some ongoing negotiations without adding or deleting any negotiation. For example, an agent may modify its negotiation strategy for a negotiation when the number of opponents in the negotiation is changed. An agent can modify its utility function according to its new expectation on negotiation outcome. An agent may delete an existing restriction between two related negotiations or generate a new restriction between two independent negotiations. In Figure 3(c), an example of these operations in a MNN is demonstrated.

### 3 Decision Making in a MNN

Because a MNN may contain more than one negotiations, and these negotiations are processed concurrently, whether to accept or reject an offer or even quit from an ongoing negotiation involves a decision making process during negotiations. An agent's decision on a single negotiation may impact its other negotiations or even the whole MNN. This section introduces an efficient procedure which can help agents to make advisable decisions for each negotiation in a MNN in order to optimize the outcome of the MNN by considering both joint utility and success rate.

### 3.1 Multi-Negotiation Influence Diagram

Suppose there are  $I$  negotiations in a MNN =  $\langle \mathbb{G}, \mathbf{R}, \mathbf{P}, \Phi \rangle$ . The decision problem in the MNN is how to make an advisable decision policy for all related negotiation in order to optimize the outcome of the MNN. A decision policy is a set of decisions that the agent makes for all negotiations in the MNN. In general, agents could have three typical decisions on an ongoing negotiation, which are (1) to *accept* the best offer from opponents, (2) to *reject* all offers and send a counter-offer and (3) to *quit* the negotiation. If a MNN contains  $I$  negotiations, the number of total decision policies for the MNN is  $I^3$ , and each policy will generate different outcomes for the MNN. In order to model the relationship between decision policies and corresponding global outcomes, we propose Multi-Negotiation Influence Diagram.

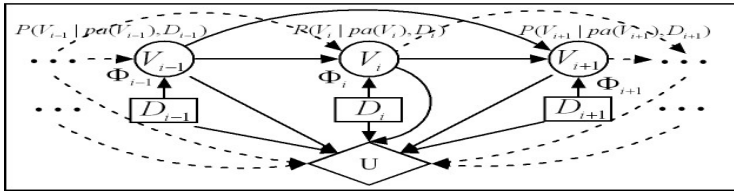


Fig. 4. A Multi-Negotiation Influence Diagram

A MNID can be defined by a six-tuple  $\langle \mathbb{G}, \mathbf{R}, \mathbf{P}, \Phi, \mathbf{D}, U \rangle$ , where  $\mathbb{G}, \mathbf{R}, \mathbf{P}, \Phi$  are same as in a MNN, and set  $\mathbf{D}$  indicates a negotiation policy and  $U$  indicates the joint utility of the MNID by considering all related negotiations.  $D_i = \{a, r, q\}$  ( $D_i \in \mathbf{D}$ ) indicates three possible decisions for each Negotiation  $V_i$  in the MNID, where  $a$  indicates *accept*,  $r$  indicates *reject*, and  $q$  indicates *quit*. A MNN can be extended to a MNID by adding a rectangular node  $D_i$  for each Negotiation  $V_i$  and one diamond node  $U$  for the whole MNN. Also, the edge from each Decision  $D_i$  to the corresponding Negotiation  $V_i$  and edges from all Decision  $D_i$  and Negotiation  $V_i$  to node  $U$  should be added. In Figure 4, a MNID is illustrated. Let  $u(\mathbf{D})$  be the joint utility of the MNID based on decisions  $\mathbf{D}$ , and  $p(\mathbf{D})$  indicates the joint success rate, and  $EU(\mathbf{D})$  indicates the expected utility, then

$$u(\mathbf{D}) = \sum_{i=1}^I u_i(D_i) \times w_i \tag{7}$$

$$p(\mathbf{D}) = \prod_{i=1}^I P(V_i|pa(V_i), D_i) \tag{8}$$

$$EU(\mathbf{D}) = p(\mathbf{D}) \times u(\mathbf{D}) \tag{9}$$

where  $w_i$  ( $\sum_{i=1}^I w_i = 1$ ) is the preference on Negotiation  $V_i$ ,  $u_i(D_i)$  is the utility of Negotiation  $V_i$  by performing Decision  $D_i$ , and  $P(V_i|pa(V_i), D_i)$  is the success rate of Negotiation  $V_i$  by considering all dependent negotiations and Decision  $D_i$ . Finally, the optimal policy for a MNID is calculated as follows:

$$\pi = \arg \max_{\mathbf{D}} (EU(\mathbf{D})) \tag{10}$$

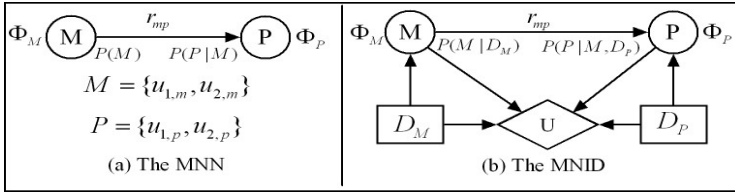


Fig. 5. The MNN and MNID for the experiment

## 4 Experiment

### 4.1 Experiment Setup

Suppose that Agent  $b$ 's global goal is to get a mortgage and to purchase a property with the mortgage, so Agent  $b$  needs to perform two negotiations. The first negotiation, (*mortgage negotiation*), is processed between Agent  $b$  and two bankers (Opponents  $o_{m1}, o_{m2}$ ) on the issues of mortgage amount and interest rate. The second negotiation, (*property negotiation*), is processed between Agent  $b$  and two real estate agents (Opponents  $o_{p1}, o_{p2}$  or Opponents  $o_{p3}, o_{p4}$ ) on the issue of the property price. It is assumed that Agent  $b$  believes that the *property negotiation* depends on the result of the *mortgage negotiation*. In Figure 5, the MNN and MNID for Agent  $b$  are displayed. Circle nodes  $M$  and  $P$  indicate the *mortgage negotiation* and the *property negotiation*, respectively. Rectangular nodes  $D_M$  and  $D_P$  are decisions on two negotiations, respectively. Diamond node  $U$  is the joint utility of the MNID. We adopt equal weighting between these two negotiations, so  $w_m = w_p = 0.5$ . Because Agent  $b$  cannot afford a property price which is higher than the mortgage amount, the restriction from *mortgage negotiation* to *property negotiation* is  $r_{mp}$ , which indicates ‘the reserved property price is the mortgage amount’. Negotiation parameters for the two negotiations are listed in Table 1 and Table 2, respectively. Because *mortgage negotiation* contains two issues, (i.e. mortgage amount and interest rate), we adopt *package deal procedure* [1] for this multi-issue negotiation and equally weight two issues. We demonstrate experimental results in two scenarios, i.e. a successful scenario, (Scenario A) and an unsuccessful scenario, (Scenario B). In Scenario A, Agent  $b$  negotiates with Opponents  $o_{m1}, o_{m2}, o_{p1}$  and  $o_{p2}$ , while in Scenario B, Agent  $b$  negotiates with Opponents  $o_{m1}, o_{m2}, o_{p3}$  and  $o_{p4}$ .

Table 1. Parameters for mortgage negotiation

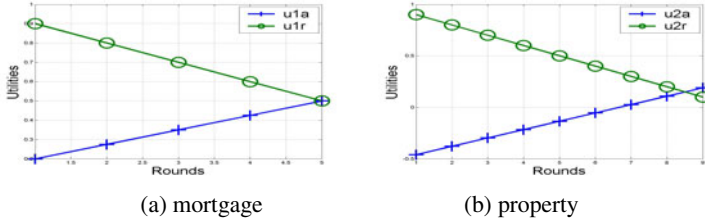
Agent	Initial Offer	Reserved Offer	Deadline
Agent $b$	(500k, 5%)	(300k, 7%)	10
Opponent $o_{m1}$	(310k, 6.9%)	(450k, 5.2%)	15
Opponent $o_{m2}$	(330k, 6.5%)	(500k, 5.5%)	9

### 4.2 Scenario A (a Successful Scenario)

In Scenario A, Agent  $b$  negotiates with Opponents  $o_{m1}$  and  $o_{m2}$  for *mortgage negotiation*, (the first negotiation) and with Opponents  $o_{p1}$  and  $o_{p2}$  for *property negotiation*,

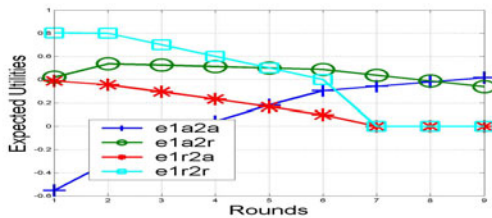
**Table 2.** Parameters for property negotiation

Agent	Initial Offer	Reserved Offer	Deadline
Agent $b$	200k	depends	12
Opponent $o_{p1}$	550k	330k	15
Opponent $o_{p2}$	500k	350k	9
Opponent $o_{p3}$	650k	450k	10
Opponent $o_{p4}$	630k	470k	11



**Fig. 6.** Negotiations using NDF approach for scenario A

(the second negotiation). Firstly, we adopt the NDF approach [4] to sequentially process *mortgage negotiation* and *property negotiation* by using the linear concession strategy, and the outcomes of the two negotiations are illustrated in Figure 6. Let letter  $a$  indicate *accept*, letter  $r$  indicate *reject* and letter  $u$  indicate utility. For instance, the legend  $u1a$  (or  $u1r$ ) indicates the utility of the first negotiation by accepting (or rejecting) opponents' offers. In Scenario A, both negotiations successfully reached an agreement by adopting the NDF negotiation model. The utility of *mortgage negotiation* is 0.5, and the utility for *property negotiation* is 0.19. Because these two negotiations are equally weighted, the overall utility is 0.35.



**Fig. 7.** Negotiations using MNN approach for scenario A

The negotiation outcomes by using the proposed MNN approach are illustrated in Figure 7. Both the *mortgage negotiation* and *property negotiation* are synchronously processed. Agent  $b$ 's reserved price in *property negotiation* is dynamically updated in each negotiation round according to the latest offer from *mortgage negotiation*. Let letter  $e$  indicate expected utility. For instance, Legend  $e1a2r$  indicates the expected utility of the MNID by accepting the best offer from opponents in *mortgage negotiation*



and rejecting all opponents' offers in *property negotiation*. The expected utility for the MNID is illustrated in Figure 7. It can be seen that before round-6, the curve  $e1r2r$  leads to the highest expected utility; from round-6 to round-8, the curve  $e1a2r$  leads to the highest expected utility; after round-8, the curve  $e1a2a$  leads to the highest expected utility. Therefore, in order to maximize the outcome of the MNID, Agent  $b$  should reject all opponents' offers in both negotiations in the first five rounds. At round-6, Agent  $b$  should accept the best offer from opponents in *mortgage negotiation* but keep on bargaining in *property negotiation* until round-8. At round-9, Agent  $b$  should accept the best offer in *property negotiation*. By adopting such a decision policy, the utility of *mortgage negotiation* is increased to 0.58 and the utility of *property negotiation* is increased to 0.26, so the global utility is increased to 0.42, which is 20% more than the result from the NDF approach.

The result of Scenario  $A$  indicates that if the global goal of related negotiations can be achieved, the proposed approach can improve the negotiation outcome through considering both joint success rate and joint utility. With comparison to sequential negotiation processes, the proposed approach can synchronously process all related negotiations and dynamically optimize the global outcome.

### 4.3 Scenario B (an Unsuccessful Scenario)

In Scenario  $B$ , Agent  $b$  negotiates with Opponents  $o_{m1}$  and  $o_{m2}$  in *mortgage negotiation* and with Opponents  $o_{p3}$  and  $o_{p4}$  in *property negotiation*. Also, we adopt the NDF approach (linear concession strategy) to sequentially process *mortgage negotiation* and *property negotiation*. The outcomes of the two negotiations are illustrated in Figure 8. In contrast to Scenario  $A$ , Agent  $b$  successfully completes *mortgage negotiation*, but fails *property negotiation*. In this case, the result of *mortgage negotiation* is meaningless or even has a negative impact by considering the global goal of related negotiations. That is because without purchasing a property, the approval of a mortgage proposal can only lead to an unnecessary cost on mortgage interest and a penalty from the bank. Therefore, if Agent  $b$  is not absolutely sure that the global goal of its related negotiations can be finally achieved, it is not efficient to process these negotiations sequentially.

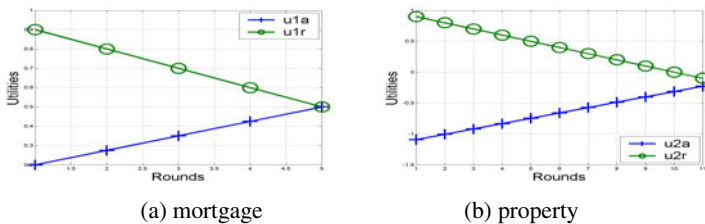


Fig. 8. Negotiations with NDF approach for scenario B

However, if we employ the proposed approach for Scenario  $B$ , the outcome will be different. In Figure 9, we illustrate the experimental results by adopting the proposed approach. In order to avoid partially reaching the global goal, Agent  $b$  can only select

policies between curves  $e1a2a$  and  $e1r2r$  (see Figure 9), which means to accept or reject both negotiations together. It can be seen that before round-8, curve  $e1r2r$  exceeds the curve  $e1a2a$ . At round-8, curve  $e1a2a$  can bring more utility to Agent  $b$  than curve  $e1r2r$ . It seems that Agent  $b$  can accept opponents' offers in both negotiations at round-8. However, because Agent  $b$  cannot purchase a property whose price is higher than the mortgage amount, so the utility of *property negotiation* must be greater than 0. At round-8, by accepting the best offer from opponents, Agent  $b$  will lose utility by 0.17, so Agent  $b$  cannot reach agreement in both negotiations at round-8. However, if Agent  $b$  stays on curve  $e1r2r$  at round-8, the expected utility will be a negative number as well in round-9. Therefore, in order to avoid any loss, Agent  $b$  can not choose neither to *accept* nor to *reject* both negotiations at round-8, but to *quit* from negotiations without achieving any agreement with any opponent. So Agent  $b$  does not need to worry about the unnecessary interest and the penalty from the bank anymore. The results of Scenario  $B$  indicate that if the global goal of related negotiations can not be achieved, then the proposed approach can help agents to avoid unnecessary losses caused by the sequential procedure.

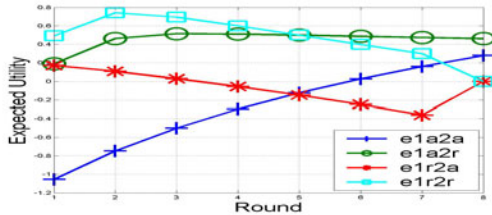


Fig. 9. Negotiations with MNN approach for scenario B

## 5 Related Work

X. Zhang and V. Lesser [10] proposed a meta-level coordination approach to solve negotiation chain problems in semi-cooperative multi-agent systems. In a complex negotiation chain scenario, agents need to negotiate concurrently in order to complete their goals on time. The order and structure that negotiations occur may impact the expected utility for both individual agents and the whole system. A pre-negotiation approach is introduced to transfer meta-level information, such as starting time, deadlines and durations of negotiation. By using this information, agents can estimate successful probability of negotiations, and model the flexibility of negotiations. However, the success probability in their work is calculated based only on a predefined time schedule, while the success rate in our work is dynamically calculated by considering an agent's possible payoffs in all related negotiations.

Proper and Tadepalli [9] proposed an assignment-based decomposition approach by employing the Markov Decision Process (MDP) to solve an optimal decision making problem in assignment decomposition between multiple collaborative agents. A centralized controller which has relevant information about the states of all agents is assumed in their approach. The approach contains two levels, where the upper level focuses on

task assignment, and the lower level focuses on task execution. The centralized controller solves the assignment problem through a searching algorithm and solves the task execution problem through coordinated reinforcement learning. The difference between our approach and their approach is that (1) we do not assume a centralized controller, and (2) we optimize the outcome of a global object through balancing its payoff and opportunity.

## 6 Conclusion

In the real world, an agent may need to process several related negotiations in order to reach a global goal. Most of state-of-the-art approaches perform these related negotiations sequentially. However, because the result of the latter negotiation is not predictable by using a sequential procedure, agents cannot optimally execute all negotiations in correct sequential order. The motivation of our approach is to solve such a problem and handle MRN concurrently. Firstly, MNN is proposed to represent MRN by considering several key features. Secondly, MNID is proposed to handle an agent's possible decisions by employing the expected utility. Lastly, through comparing expected utilities between all possible policies, an optimal policy is generated to optimize the outcome of MRN. The experimental results indicate that the proposed approach improved an agent's global utility of MRN in a successful end scenario, and avoid unnecessary losses for the agent in an unsuccessful end scenario.

## References

1. Fatima, S., Wooldridge, M., Jennings, N.: An Agenda-Based Framework for Multi-Issue Negotiation. *Artificial Intelligence* 152(1), 1–45 (2004)
2. Li, C., Giampapa, J., Sycara, K.: Bilateral Negotiation Decisions with Uncertain Dynamic Outside Options. *IEEE Trans. on Systems, Man, and Cybernetics, Part C* 36(1), 31–44 (2006)
3. Ren, F., Zhang, M., Sim, K.: Adaptive Conceding Strategies for Automated Trading Agents in Dynamic, Open Markets. *Decision Support Systems* 46(3), 704–716 (2009)
4. Faratin, P., Sierra, C., Jennings, N.: Negotiation Decision Functions for Autonomous Agents. *J. of Robotics and Autonomous Systems* 24(3-4), 159–182 (1998)
5. Rubinstein, A.: Perfect Equilibrium in a Bargaining Model. *Econometrica* 50(1), 97–109 (1982)
6. Fatima, S., Wooldridge, M., Jennings, N.: An Analysis of Feasible Solutions for Multi-Issue Negotiation Involving Nonlinear Utility Functions. In: *Proc. of 8th Int. Conf. on AAMAS 2009*, pp. 1041–1048 (2009)
7. Brzostowski, J., Kowalczyk, R.: On Possibilistic Case-Based Reasoning for Selecting Partners for Multi-Attribute Agent Negotiation. In: *Proc. of 4th Int. Conf. on AAMAS 2005*, pp. 273–279 (2005)
8. Hemaissia, M., Seghrouchni, A., Labreuche, C., Mattioli, J.: A Multilateral Multi-Issue Negotiation Protocol. In: *Proc. of 6th Int. Conf. on AAMAS 2007*, pp. 939–946 (2007)
9. Proper, S., Tadepalli, P.: Solving Multiagent Assignment Markov Decision Processes. In: *Proc. of 8th Int. Conf. on AAMAS 2009*, pp. 681–688 (2009)
10. Zhang, X., Lesser, V.: Meta-Level Coordination for Solving Negotiation Chains in Semi-Cooperative Multi-Agent Systems. In: *Proc. of 6th Int. Conf. on AAMAS 2007*, pp. 50–57 (2007)
11. Jensen, F., Nielsen, T.: *Bayesian Networks and Decision Graphs*, 2nd edn. Springer, Heidelberg (2001)

# Reasoning Activity for Smart Homes Using a Lattice-Based Evidential Structure

Jing Liao, Yaxin Bi, and Chris Nugent

Computer Science Research Institute  
School of Computing & Mathematics  
University of Ulster  
Northern Ireland, UK

`liao-jl@email.ulster.ac.uk, {y.bi,cd.nugent}@ulster.ac.uk`

**Abstract.** This paper explores a revised evidential lattice structure designed for the purposes of activity recognition within Smart Homes. The proposed structure consists of three layers, an object layer, a context layer and an activity layer. These layers can be used to combine the mass functions derived from sensors along with sensor context and can subsequently be used to infer activities. We present the details of configuring the activity recognition process and perform an analysis on the relationship between the number of sensors and the number of layers. We also present the details of an empirical study on two public data sets. The results from this work has demonstrated that the proposed method is capable of correctly detecting activities with a high degree of accuracy (84.27%) with a dataset from MIT [4] and 82.49% with a dataset from the University of Amsterdam[10].

**Keywords:** Activity recognition; Smart Homes; sensor fusion; reasoning under uncertainty; Dempster-Shafer theory of evidence.

## 1 Introduction

The population of the UK is ageing [1]. Population ageing will bring a new set of challenges to society such as health and social care. One possible solution is through the introduction of remote monitoring technology which can offer a means of providing independent living and subsequently increase perceptions of improved social care. Through automatically inferring human activity caregivers and healthcare professionals can monitor the health and behavioral status of those within their own homes and provide an improved service and levels of care support.

In terms of activity recognition, there are many types of sensors deployed within smart environments to assist with the recognition process. This requires the use of data fusion techniques to combine data from multiple sensors, and relevant information from associated databases, to achieve improved accuracies and more specific inferences than could be achieved by the use of a single sensor alone [2]. Among data fusion techniques, the Dempster-Shafer (D-S) theory is recognized as an effective approach which can be used to handle uncertainty and to combine sensor data[3].

Examples of applying the D-S theory of evidence to activity recognition in smart homes have been previously reported in [5,6,9]. In [5] Hong has created a general ontology for managing activities of daily living in a smart environment. The study did not however, address a quantitative analysis of the uncertainty existing in Smart Homes and along with a practical way to calculate the discounting value. In [9] Hyun has proved that the D-S theory can provide an efficient way to incorporate and reduce the impact of uncertainty. This research only focuses on reducing the uncertainty interval from the belief value to the plausibility value instead of improving the belief value.

In our previous work [6] we have created a framework based on the D-S theory to manage the uncertainty which may exist within the activity recognition process within smart homes. Within that work we have identified two kinds of uncertainty sources, one is from the hardware, called hardware uncertainty. This type of uncertainty originates from errors relating to hardware such as sensor errors (for example one of the sensors is broken, the sensor's battery is exhausted) and a transmission error between the receiver and transmitter [8]. The other kind of uncertainty is related to human activity, which is referred to as context uncertainty. For example, a person may engage in a multitask based activity. Here the duration of each activity is not the same and the activity sequence is complex. All of these factors will cause uncertainty to some extent. According to the process of generating the uncertainty two discounting values have been used to represent the two kinds of uncertainty. Using Dempster's combination rule we fuse the evidence in the object layer which will cause the loss of some information from the indirect observation such as the situation of two objects involved together.

In an effort to enhance the performance of our previous work and in an attempt to manage domain information we have proposed a lattice-based structure with three types of layers: an object layer, a context layer and an activity layer. In addition we have developed a belief propagation algorithm for the structure. In particular the context layer could consist of a varying number of sub-layers so as to accommodate complex multitask activities. To maximize the effect of sensor combination we also investigate relationship between the numbers of sensors and the numbers of layers within the lattice structure through examination of their performance in the analysis of two datasets.

We have used two public datasets to evaluate the performance of the belief propagation algorithm defined for our revised lattice structure. Based on the indirect record for the instance of an activity in a dataset the evaluation was performed using the leave-one-out cross-validation, i.e. 13 days of data were used for training and one day of data was used for testing.

## 2 Lattice Structure

As previously introduced in [6], sensor combination can be based on the evidence obtained from single sensor objects which means that the system does not consider historical data and the activity pattern recorded through indirect observations. In order

to combine the information from the indirect records and incorporate the uncertainty existing in Smart Homes we have developed a lattice based structure which can be used to fuse multiple pieces of evidence in a composite context for activity recognition.

### 2.1 General Lattice Structure for Activity Recognition

The lattice based structure is composed of three layers, namely object, context and activity layers. The object layer consists of a number of sensors which are installed on various objects and appliances within the smart home. The context layer comprises combinations of sensors, which are recursively constructed on the basis of the sensors involved. The activity layer contains the activities being considered. The advantage of the proposed structure is that it supports the system to use historical data as *a priori* knowledge to adjust the belief and mass values derived from the sensors and the sensor context and finally to improve the accuracy of activity recognition. The general architecture of the lattice structure for activity recognition is shown in Figure 1.

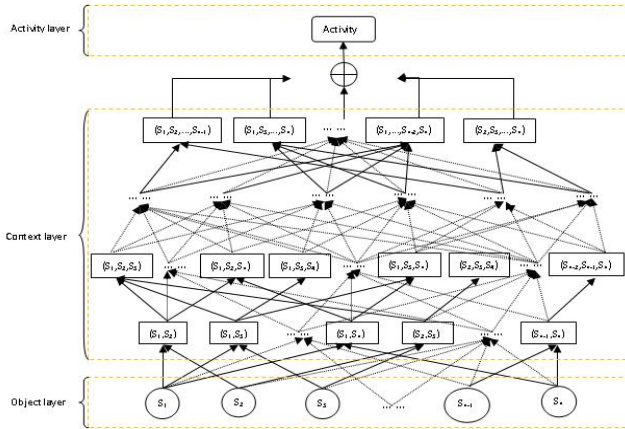


Fig. 1. The general lattice structure for activity recognition

In Fig. 1, circular nodes represent sensors and related objects, composite context nodes are represented by a rectangle and a rounded rectangular node represents an activity.

Given that there are  $n$  sensors involved in one activity such as preparing the dinner, the lattice structure can be used to reason about the activity being performed. The object layer contains  $n$  nodes representing  $n$  sensors and the related objects are denoted by  $(S_1, S_2, \dots, S_n)$ . The context layer is composed of  $n-2$  sub-layers, namely context layer 1, context layer 2, until context layer  $n-2$ . The context layer 1 consists of  $C_n^2$  context nodes. Each node contains two elements generated from two objects in the object layer, i.e. there is a mapping from the two objects to the node. The context layer 2 is composed of  $C_n^3$  context nodes, each of which is generated from three

context nodes in the context layer 1. Repeating the above process, we can recursively construct the context layer  $n-2$  containing  $C_n^{n-1}$  context nodes which are generated from context nodes in the context layer  $n-3$ . The activity layer will present the activity involved.

## 2.2 The Algorithm for Belief Propagation

Using the lattice structure the activity inference is based on 3 layers – object layer, context layer and activity layer. In the object layer the degree of belief of each sensor records can be represented by a mass function with the two conditions[3]:

- (1)  $m(\emptyset) = 0$      $\emptyset$ : the empty set
- (2)  $\sum_{A \subseteq \Theta} m(A) = 1$      $A$ : a subst of  $\Theta$

The mass value of each object node can be obtained by the sensor status (active or inactive) and the discounting value can be defined on the basis of sensor manufacture statistics. Consider sensor node  $S_i$  as an example, given that the  $S_i$  sensor's status is active, hence the initial mass value of sensor node  $S_i$  is 1. Through the process of discounting a revised mass value of  $S_i$  can be calculated by formula (1)[3].

$$m^d(A) = \begin{cases} (1-r)m(A) & A \subset \Theta \\ r + (1-r)m(\Theta) & A = \Theta \end{cases} \quad (1)$$

where  $r$  is the discounting rate with a value range of 0 to 1. When  $r$  is equal to 0, it means the source is completely reliable. In contrast if  $r$  is equal to 1, it means that the source is absolutely unreliable.

The whole procedure of calculating the discounting value and the mass value of the object node is reported in [6].

As previously introduced the context layer encompasses  $n-2$  sub-layers, the strength of mapping among context layers is different. The mass value of each context node in context layer  $i$  ( $i=2, 3, \dots, n-2$ ) is translated from the context layer  $i-1$ . Based on the layer topology, we define a weighting factor to quantify the strength of mappings. Through accounting for the frequencies of occurrence of each context node in the context layer, the weighting factor ( $W_1, W_2, \dots, W_{n-i+1}$ ) from the context node in context layer  $i-1$  to the relative context node in context layer  $i$  can be quantified by the frequency of each context node. Taking context node ( $S_1, S_2, \dots, S_i$ ) in context layer  $i$  as an example the algorithm for computing the weighting factors ( $W_1, W_2, \dots, W_{n-i+1}$ ) of context node ( $S_1, S_2, \dots, S_i$ ) in context layer  $i$  ( $i=1, 2, \dots, n-2$ ) is followed.

```

Program general weighting factor calculation (output)
{Assuming the number of event ( $S_1, S_2, \dots, S_{i-1}$ ), ( $S_1, S_3, \dots, S_i$ ),  $\dots$ , ( $S_{n-i+1}, S_{n-i+2}, \dots, S_n$ ) triggered  $N_1, N_2, \dots, N_{n-i+1}$  times;
the number of event ( $S_1, S_2, \dots, S_i$ ) triggered time is  $S_i$ ;}
Var  $N_1, N_2, \dots, N_{n-i+1}$ 
 $S_i$ 
 $W'_1, W'_2, \dots, W'_{n-i+1}$ 
 $W_1, W_2, \dots, W_{n-i+1}$ 

```

```

Begin
  j=1;
  repeat
    j=j+1;
    W'_j = S_1 / N_j;
  until j= n-i+1;
  W_1 = W'_1 / \sum_{j=1}^{n-i+1} W'_j ;
  W_2 = W'_2 / \sum_{j=1}^{n-i+1} W'_j ;
  ...
  W_{n-i+1} = W'_{n-i+1} / \sum_{j=1}^{n-i+1} W'_j;
end

```

Since the Dempster's rule of combination can aggregate a group of evidence obtained from single or multiple sources, in the activity layer we use the Dempster's rule to combine the belief value of each context node in context layer  $n-2$ . The formula for fusing two pieces of evidence is shown below[3].

$$m_1 \oplus m_2(C) = \begin{cases} k \sum_{A_i \cap B_j = C}^{i,j} m_1(A_i) * m_2(B_j) & \text{if } A_i \cap B_j \neq \emptyset \\ 0 & \text{if } A_i \cap B_j = \emptyset \end{cases} \quad (2)$$

Where  $k$  is the conflict factor between the two pieces of evidence  $A$  and  $B$  that can be obtained by formula (3) below[3].

$$k = (\sum_{A_i \cap B_j \neq \emptyset}^{i,j} m_1(A_i) * m_2(B_j))^{-1} \quad (3)$$

### 3 Case Study

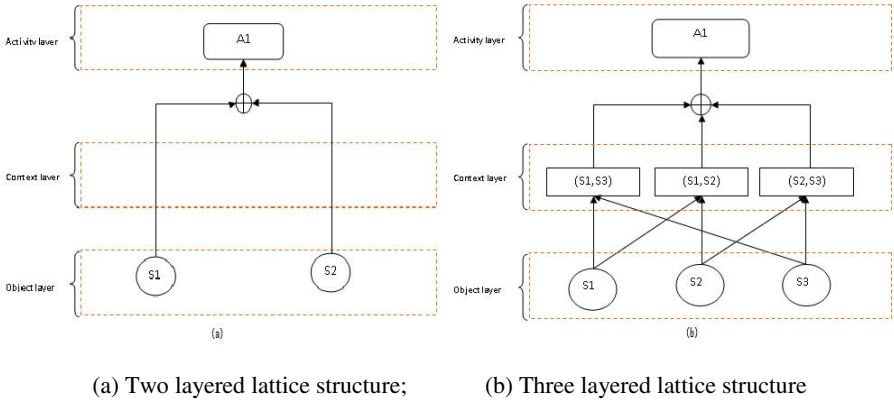
To help illustrate how the combination structure works we have used the activity of toileting as an example. In this example consider that an ageing person is going to use the toilet in which there are 5 sensors in the bathroom environment used to detect their behavior. The sensors include bathroom light, toilet flash, bathroom sink tap (cold), bathroom sink tap (hot) and bathroom cabinet. This is not an exclusive list and in other scenarios there may be more than 5 sensors involved. Nevertheless, here we have only selected the key sensors for illustrative purposes.

Based on the five sensors triggered, there are four groups of lattice-based structures that can be selected. There are two sensors with two layers, three sensors with three layers, four sensors with four layers and five sensors with five layers. The following sections will analyze how the accuracy of activity recognition is related to each group of the lattice structures.

#### 3.1 Two Layered Lattice Structure

As shown in Fig. 2(a), the two layered lattice structure only contains two nodes in the object layer and the toileting and non-toileting activities in the activity layer. It does not generate context nodes in the context layer.



**Fig. 2.**

In the case study there are five sensors triggered, however, only two sensors have been selected. Thus we have 10 sets of sensor combinations, which can be used for toileting activity analysis. The 10 sets of sensors are listed in Table 1. We used the dataset from MIT lab [7] to analyze the toileting activity recognition process. The results from this analysis are presented in Table 1.

In Table 1, F stands for the toilet flush, L represents the bathroom light, C refers to the bathroom cabinet, and bathroom sink tap (hot) and sink tap (cold) are denoted by TH and TC, respectively. To evaluate the performance of this structure we compared the system report with the activity sequence that the subject reported after the experiment. The latter is considered as our ground truth. We use the true positive (TP) metric to present a correct claim of the toileting activity as having occurred and the false positive (FP) metric to represent an incorrect claim. The false negative (FN) metric represents that the activity occurred but was not identified. The evaluation methods used were precision, recall and F-Measure [11].

From Table 1, it is evident that the (flush, light) set and (flush, tap(hot)) have the highest accuracy levels of 90% for toileting activity recognition. The mean precision is 67.61% and the mean recall is 89.65%.

### 3.2 Three Layered Lattice Structure

Fig. 2(b) presents the three layered lattice structure that may be used to combine information from 3 sensors or objects.

As shown in Fig.2 (b), the object layer contains three object/sensor nodes and there is only one context layer generated. The context layer is composed of three context nodes. The activity layer still only contains the toileting activity and non-toileting activity as shown in Fig. 2 (a).

In the toileting case study, the three layered lattice structure only selects three sensors for combination although there are ten sets of sensor combinations. The 10 sets of sensors and the toileting recognition accuracy results are presented in Table 2.

From Table 2 we can see that the highest precision and recall is 90.59% obtained from the sensor set (flush, cabinet, tap (hot)). The mean precision for activity recognition is 85.73% and the mean recall is 80.35%.

**Table 1.** The recognition results of the two layered structure<sup>1</sup>

Sensor set	TP	FP	FN	Precision (%)	Recall (%)	F-Measure (%)
F,L	72	3	13	96.00%	84.71%	90.00%
F,C	69	3	16	95.83%	81.18%	87.90%
L,C	77	32	8	70.64%	90.59%	79.38%
L,TH	84	80	1	51.22%	98.82%	67.47%
L,TC	68	52	17	56.67%	80.00%	66.34%
C,TC	68	52	17	56.67%	80.00%	66.34%
C,TH	84	80	1	51.22%	98.82%	67.47%
F, TH	84	81	1	50.91%	98.82%	67.20%
F,TC	72	3	13	96.00%	84.71%	90.00%
TC,TH	84	81	1	50.91%	98.82%	67.20%
MEAN				67.61%	89.65%	74.93%

**Table 2.** The recognition results of the three layered lattice structure<sup>1</sup>

Sensor set	TP	FP	FN	Precision (%)	Recall (%)	F-Measure (%)
F,L,C	62	2	23	96.88%	72.94%	83.22%
F,L,TH	66	2	19	97.06%	77.65%	86.27%
F,L,TC	53	0	32	100.00%	62.35%	76.81%
F,C,TH	77	8	8	90.59%	90.59%	90.59%
F,C,TC	67	3	18	95.71%	78.82%	86.45%
L,C,TH	76	8	9	90.48%	89.41%	89.94%
L,C,TC	60	4	25	93.75%	70.59%	80.54%
F,TC,TH	76	46	9	62.30%	89.41%	73.43%
L,TC,TH	69	35	16	66.35%	81.18%	73.02%
C,TC,TH	77	43	8	64.17%	90.59%	75.12%
MEAN				85.73%	80.35%	81.54%

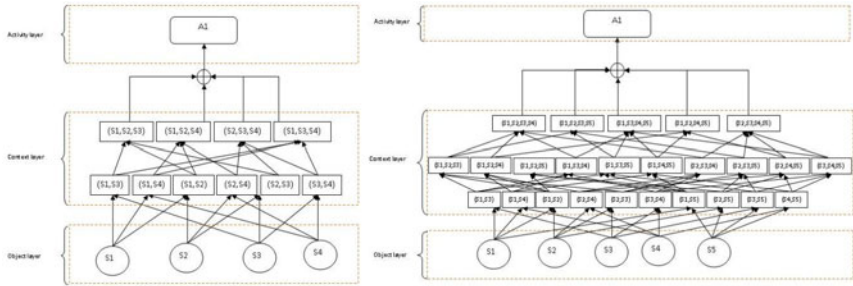
### 3.3 Four Layered Lattice Structure

More than four sensors activated can be chosen to adapt a four layered lattice structure. As shown in Fig.3 (a) the object layer contains four object/sensor nodes and the context layer encompasses two sub layers, denoted by context layer 1 and context layer 2, respectively. The activity layer includes the activity involved.

Assuming there is only one type of activity i.e. toileting activity, we use the same scenario as previously introduced to study the performance of the four layered lattice structure. This will generate five combination sets based on the former scenario and the recognition results are shown in Table 3.

From Table 3 we can see that the (flush,light,cabinet, tap (hot) ) combination set achieves the highest precision 100% and 81.18% recall. The mean precision of the five combination set is 91.52% and the mean recall of the combination set for toileting activity recognition is 75.29%.

<sup>1</sup> F=flush, L=light, C=Cabinet, TH=tap (hot), TC=tap (cold), TP=true positive, FP=false positive, FN=false negative, PR=precision, RE=recall, F-M= F-measure.



(a) Four layered lattice structure;

(b) five layered lattice structure

Fig. 3.

Table 3. The recognition results of the four layered lattice structure

Sensor set	TP	FP	FN	Precision (%)	Recall (%)	F-Measure (%)
F,L,C,TH	69	0	16	100.00%	81.18%	89.61%
F,L,C,TC	61	1	24	98.39%	71.76%	82.99%
F,C,TC,TH	70	3	15	95.89%	82.35%	88.61%
F,L,TH,TC	54	4	31	93.10%	63.53%	75.52%
L,C,TC,TH	66	28	19	70.21%	77.65%	73.74%
MEAN				91.52%	75.29%	82.09%

### 3.4 Five Layered Lattice Structure

From the toileting scenario we know that there are only five sensors which are triggered, according to the general lattice structure previously introduced the largest layer of lattice structure being generated is five layers structure. The five layers structure is shown in Fig.3 (b).

As shown in Fig.3 (b) the object layer includes five object/sensor nodes and the context layer consists of three sub context layers. The structure has one activity layer. For the toileting scenario, the five object nodes are the toileting flush, bathroom light, bathroom sink tap (hot) and sink (cold), and bathroom cabinet. There are only five active sensors and there will be only one combination set for a five layered lattice structure. Based on the MIT dataset, the system based on this structure identified 71 toileting activities correctly (TP), incorrectly identified 3 toileting activities (FP) and missed 14 toileting activities (FN). The total precision we obtained was 95.95% with a recall of 83.53%.

### 3.5 Results Comparison

As previously shown we have four groups of lattice structures for the purposes of toilet activity recognition given that there are five active sensors. We compare the mean accuracy of four groups of lattice structures in Table 4.

**Table 4.** The mean accuracy of the four types of lattice structure for the toileting activity

Sensor number	layers	Precision (%)	Recall (%)	F-Measure (%)
2	1	67.61%	89.65%	74.93%
3	2	85.73%	80.35%	81.54%
4	3	91.52%	75.29%	82.09%
5	4	95.95%	83.53%	89.31%

As shown in Table 4 with the increase of the number of sensors in the process of reasoning the toileting activity, the accuracy increase from 74.93% to 89.31% which means that the more layers we adapted and the more sensors we use, the higher the accuracy we can achieve.

The layered structure can distinguish between the minor differences from the sensor information. For example, based on the same dataset the two layered structure cannot separate the difference between the combination set (light, tap (cold)) and the combination set (cabinet, tap (cold)). As shown in Table 1 the two combination sets all obtain 56.67% precision and 80.00% recall. Nevertheless, under the three layered structure, the system separated the difference of the two combination set, with the precision of the combination set ( light, tap (cold) , tap (hot) ) being 66.35% and the precision of the combination set (cabinet, tap (cold) , tap (hot) ) being 64.17%.

From the above evaluations we can appreciate that the best lattice structure for the toileting activity recognition is based on the approach which adapts the largest layered structure and uses the largest numbers of sensors for the analysis.

## 4 Analysis of Results

As previously reported the D-S theory with lattice structure can be used to infer activities within Smart Homes. To further evaluate the performance of the lattice structure with the D-S theory of evidence, we used two publically available datasets for testing the system's performance. The first dataset was from the AI lab of MIT [4] and the second dataset was from the Intelligent Systems Lab of the University of Amsterdam (denoted by UoA) [10].

Inferring the activity using the D-S theory starts from the definition of the frame of discernment. In this scenario we can define 4 frames of discernment as shown in Table 5. The second step is to apply the discounting value to each sensor and object node. Through the application of the discounting value the system can incorporate the uncertainty within the smart home and in human activities. Following the discounting process the next step is to perform the translating and propagation from the object layer to the context layer. In this process we calculate the weight factors in two context layers as previously introduced. The final step is to combine the evidence in the activity nodes using the orthogonal sum.

The experiments were conducted within the MATLAB 8.5 environment. The method used to evaluate the accuracy of the activity recognition algorithm was the leave-one out cross-validation method. The metrics used to evaluate performance were Precision, Recall and F-measure. The following sections present the analysis of the results.

#### 4.1 Analysis of Results Based on the MIT Dataset

In MITs' experiments [7], 77 contact sensors were installed in a single-person's apartment to collect data about a resident's activity for a period of two weeks [4]. These sensors were attached to objects such as drawers, refrigerators, containers, etc. to record activation-deactivation events when the subject carried out daily activities [7]. We chose the data collected for the subject 1 as our investigative dataset. In this subject's apartment 77 sensors were installed.

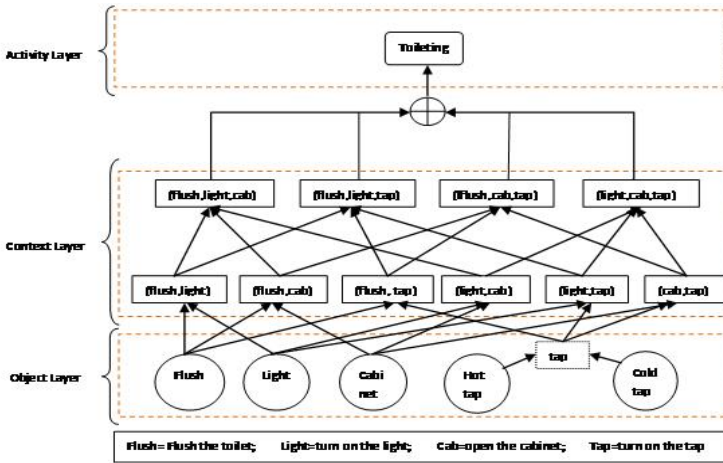


Fig. 4. The layer topology of toileting activity recognition

Based on the indirect observation of ADLs (Activity of Daily living) in the MIT dataset, we identified 5 key sensors related to the toileting activity. Two of them were found to have the same influence on the toileting activity. So we adapted the four layer lattice structure to combine the sensor data. The whole lattice structure of the toileting activity recognition using the MIT dataset is shown in Fig.4.

Table 5. Examples of frames of discernment

Name	Type	Location	Frame of discernment
sflush	Sensor	Object layer	{sflush,-sflush}
flush	Object	Object layer	{flush,-flush}
(flush,light)	Context	Context layer 1	{(flush,light), -(flush,light)}
(flush,light,cab)	Context	Context layer 2	{(flush,light,cab), -(flush,light,cab)}

Based on the structure of the layers, we can define the frames of discernment. The frame of discernment can embody all possible node elements of interest, denoted by  $\Theta$ . Every node in the layers such as sensor node, object node, context node and activity node can all be represented by one frame of discernment. Table 5 is an example of frames of discernment defined on the basis of the lattice structure in the case of recognizing the toileting activity.

Table 6 presents the results of the four layered structure based activity recognition. There are 85 toileting activities which took place on the two weeks experiments. Our system identified 93 of these toileting activities during the experiment. In addition we validated both the identified activity and the indirect records, the activity start time and duration. The 75 of the identified activities were correctly recognized and 18 activities were incorrectly reported.

**Table 6.** Results of toileting activity recognition based on two public dataset

Data set	TP	FP	FN	Precision (%)	Recall (%)	F-Measure (%)
MIT[4]	75	18	10	80.65%	88.24%	84.27%
UoA[10]	106	8	37	92.98%	74.13%	82.49%

In comparison with a Naïve Bayesian network used for analysing the same data [4], our four layer structure provided a result of 80.65% and was better than the 61.2% reported by [4]. In addition the recall 88.24% is comparable with the value of 83.5% reported in [4]. Compared with the results of 69.4% of our original method reported in [6], this new approach significantly improved the accuracy of activity recognition.

## 4.2 Analysis of Results Based on UoA Dataset

In the experimental environment [10], 14 state-change sensors were installed in a single-person apartment to collect data about resident's activities for a period of four weeks [10]. In this experiment only two sensors were configured for the purpose of recognizing toileting activity- toileting *flush* sensor and bathroom *light* sensor. The lattice structure is composed of two layers. Based on this data we examined the performance of our approach in reasoning the toileting activity. The total results from these experiments are presented in Table 6.

As shown in Table 6 our method can correctly identify 92.98% of the toileting activities. Of the toileting activity instances that occurred, this method correctly detected 74.13% of them. Compared with the hidden Markov model [10] used with the same dataset, the total accuracy of our method offers an improvement of 2.34% in accuracy, i.e. from 83.9% in [10] to 86.24%.

Overall, based on the analysis of the public data sets, our experimental results demonstrate the advantage of the revised lattice structure for the toileting activity.

## 5 Conclusions

In this paper we have introduced an improved approach to infer human activities within the smart home. With this improved approach we have built a lattice-based structure and increased the inference layers to improve the system performance. It has been shown that the use of the weight factor for the system can incorporate the context uncertainty in the information being used to recognize the activity. The results have been based on an analysis of a component of two publically available datasets

and have shown that the improved approach has proven itself to be a solid approach to recognize human activity in a smart home environment in comparison to previously developed and tested approaches.

On the other hand the more layers we used the more complex the program is hence requiring more computing time. Given that one complex activity involving many sensors such as making dinner will trigger more than 10 sensors, the best lattice structure for this activity analysis may adapt more than 10 layers of the lattice structure, thus the nodes in context layers will exponentially expand which will dramatically increase the computational complexity. In this situation we need to find out the best number of layers for activity recognition which can obtain a higher accuracy but with less computing time. The next step in our work is to dynamically adjust the number of layers to obtain the best layered structure for activity recognition and investigate how the dependence between nodes in context layers affects the accuracy of activity recognition.

## References

1. National Statistics: 2008-based national population projections (October 2009)
2. Hall, D.L., Llinas, J.: An introduction to multisensor data fusion. *Proceedings of the IEEE* 85(1), 6–23 (1997)
3. Shafer, G.: *A mathematical theory of evidence*. Princeton University Press, Princeton (1976)
4. Tapia, E.M.: *Activity recognition in the home setting using simple and ubiquitous sensors*. M.S. thesis, School of Architecture and Planning, Massachusetts Institute of Technology, USA (2003)
5. Hong, X., Nugent, C., Mulvenna, M., McClean, S., Scotney, B., Devlin, S.: Evidential fusion of sensor data for activity recognition in smart homes. *Pervasive and Mobile Computing* 5, 236–252 (2009)
6. Liao, J., Bi, Y.X., Nugent, C.: Evidence fusion for activity recognition using the Dempster-Shafer theory of evidence. In: *IEEE International Conference on Information Technology and Applications in Biomedicine*, Cyprus (2009)
7. Ferscha, A., Mattern, F.: Activity recognition in the home setting using simple and ubiquitous sensors. In: *Proceeding of pervasive 2004*, Berlin, Heidelberg, pp. 158–175 (2004)
8. Philipose, M., Fishkin, K.P., Perkowitz, M.: Inferring activities from interactions with objects, pp. 50–57. *IEEE CS and IEEE ComSoc*, Los Alamitos (2004)
9. Hyun, Lee, Choi, J.S., Elmasri, R.: Sensor Data Fusion Using DS<sub>m</sub> Theory for Activity Recognition under Uncertainty in Home-Based Care. In: *Conference on Advanced Information Networking and Applications*, USA, pp. 517–524 (2009)
10. Van Kasteren, T.L.M., Noulas, A.K., Englebienne, G., Kröse, B.J.A.: Accurate Activity Recognition in a Home Setting. In: *ACM Tenth International Conference on Ubiquitous Computing (UbiComp 2008)*, Seoul, South Korea (2008)
11. Bernardo, J.M., Smith, A.F.M.: *Bayesian Theory*. Wiley, England (2000)

# Knowledge Modelling to Support Inquiry Learning Tasks

Annika Wolff, Paul Mulholland, Zdenek Zdrahal, and Miroslav Blasko

Knowledge Media Institute and Centre for Research in Computing,  
The Open University, Milton Keynes, MK7 6AA, UK  
{A.L.Wolff, P.Mulholland, Z.Zdrahal}@open.ac.uk,  
Blasko@labe.felk.cvut.cz

**Abstract.** In this paper we describe the SILVER toolkit, which is designed for tasks in which a user learns by analysing and interpreting a set of resources. The user categorises each resource according to the set of properties that they identify as being applicable to it. Due to the large amount of data generated by this type of task, the user may find it hard to identify patterns in their classification and tagging, to recognise their own inconsistencies or make comparisons between themselves and others. In the first SILVER task described, the ID3 decision tree algorithm is applied to the user's data to identify patterns and generate different types of feedback. Principles of spatial hypertext are used to produce an interactive visualization of the summarized data. As the user interacts with the resources, they can see their progress and changing perspective on the task. In the second SILVER task described, a conceptual model is used to provide explanations of the model underlying the user's classification of resources.

**Keywords:** Spatial hypertext, inquiry learning, ID3 decision trees, categorization and organization.

## 1 Introduction

As computers become available in most classrooms and homes, teachers are keen to exploit the learning potential of the large amount of multimedia content which is now available. The SILVER project aims to provide tools to facilitate learning from collections of resources. The developed toolkit is appropriate for concept-learning tasks where a user browses a collection of objects, in this case images, discriminating between them based on a set of attributes that can be ascribed to the resource and which provide some evidence related to the concept to be learned.

The attributes are presented as a set of tags, with values, that the user can choose to apply. Taking their tagging into consideration, the user then classifies the resource according to the chosen dimensions of the task. Unfamiliarity with the domain may lead to the inconsistent application of tags and classification of resources, particularly in the earlier stages of the task. As the user progresses through the task, their knowledge of the domain should increase and their judgements should become more consistent. However, without feedback there is the possibility that they are simply



reinforcing incorrect patterns between tags and classifications. Furthermore, as the task progresses and the amount of data available continues to rise, it becomes more difficult for the user to perceive patterns in their input. The SILVER toolkit addresses these issues by using machine-learning techniques to analyse the user's data and provide succinct and useful feedback to assist them in making rapid assessments of their progress. Feedback can be used to draw a user's attention to inconsistencies in their own data, or to visually compare their data against that of another user, highlighting important differences. The user is able to make corrections and see immediately what impact this has.

The toolkit has three components. The first is the SILVER-RE (reasoning engine), processes the task data and generates feedback. The second is the Magic Studio task interface, for interacting with task materials and visualizing feedback. The third is a java plug-in for visualizing some additional functionality from the SILVER-RE.

This paper first looks at some related work, then examines the kinds of tasks supported by the SILVER toolkit. Then we discuss how these have been developed into a module for learning about sustainable development and a little about some in-class assessment of this module. Next, we look at how the SILVER-RE generates feedback based on the current state of the task and explore how principles of spatial hypertext are used to visually represent this feedback to the user. Finally, we examine how a new task has been developed for learning about leadership styles and how this is supported by the addition of a conceptual model.

## 2 Related Work

In a spatial hypertext, the visual organisation and attributes of a set of objects is seen as an important information source to the user. The objects used within a spatial hypertext are commonly surrogates [1], which represent and provide access to a more complex resource. Visual representations exploit people's inherent ability to quickly process and comprehend a visual scene. As such, they provide a good medium for the rapid sharing of information, since users can easily understand one another's visual representations of a task even if not structured in exactly the same way. Visual similarity and co-location of items are two properties that are commonly utilised within spatial hypertext systems to provide information about a set of objects [2]. VIKI [3] is an early example of a spatial hypertext system which used the layout of objects to convey meaning and to support users in information triage. The Virtual Knowledge Builder [4] uses both layout and visual attributes to support a user in the incremental creation and interpretation of visual knowledge representations. One of the main goals of Knowledge Building is to produce external representations that can be understood and worked upon by more than one person [5]. On a similar basis, VITE [6] helps users to create and use visual representations of structured data in problem solving tasks. Garnet [7] uses spatial hypertext to support the identification of patterns across a set of retrieved content from a digital library.

The visual representations provided in SILVER consist of a spatially organised set of surrogates, each being a thumbnail image which represents and provides access to a single resource. The surrogates are organised according to the user's classification

of the resource, which in turn is based on their tagging. Additional information is provided by way of colour-coded markers attached to the thumbnails. Principles of spatial hypertext assist a user in identifying patterns in their data, comparing these patterns with other users and spotting inconsistencies and possible errors. The interface display changes as the user interacts with it, and can:

- show how resources have been tagged and classified.
- use machine-learning techniques to identify and show patterns between the tagging and the classification of resources.
- draw a user's attention to inconsistencies in their tagging and classification, allowing them to reflect on their choices.
- enable direct manipulation of objects on the screen, so that the user can make changes and see immediately what effect these have.
- allow comparisons between users, highlighting important differences.
- use surrogates to represent any kind of resource to which tags and classifications can be applied, including different media types and also people – e.g. for visualizing patterns and comparing people's views across a group.

### 3 Silver Tasks

The SILVER toolkit supports tasks in which a user is required to look for evidence amongst a set of resources which can be used to answer a question. In one task, the SILVER toolkit has been used to answer questions about sustainable development e.g. “what factors contribute to the environmental, economic or social sustainability of a building?”. The sustainable development task will be used within this paper to illustrate the capabilities of the SILVER toolkit for summarizing and visualizing a user's complex data set.

SILVER tasks are presented via the Magic Studio interface. Magic Studio can be used to construct task modules that can then be configured in different ways, according to the specific requirements of the subject and of the individual teacher. The sustainable development task is presented via several stages, which each follow a general pattern of tagging, followed by the visualization of some summary information. In this paper, we focus on one pattern of tagging (Tag Buildings) and visualization (View your Tagging), which are described next.

In the ‘tag buildings’ task, the user expresses what they have found in a resource by assigning attributes from a pre-specified list. In the sustainable development task, attributes included building materials such as glass, brick, wood and concrete. Giving such a list provides structure to the task and allows the user's input to be interpreted by the SILVER-RE. Figure 1 shows an example of a tagging task. The building they are tagging is on the left hand side and the available tags are on the right. The right-hand pane contains all the available images for the task. The user can obtain more information about a building either through the more information button, which might link to text, a movie or an audio file, or else through hotspots associated with parts of the image. A hotspot might link to another image, possibly revealing more detail

Tag buildings

Help Off

**Buildings Keywords**

- Community facility
- Pollution
- Grass
- Trees and flowers
- River
- Place for trade business or tourism
- Industrial building
- Concrete
- Thatched roof
- Stone
- Brick
- Glass
- Wood

**Jobs**

- None
- Some
- Many

**Housing**

- None
- Low-density
- High-density

Buildings

1 What is sustainable development? 2 Tag buildings 3 View your tagging EXIT

Produced by Bridgeman Education, Lexara Ltd and The Knowledge Media Institute. Funded by the Department for Innovation, Universities and Skills (DIUS)

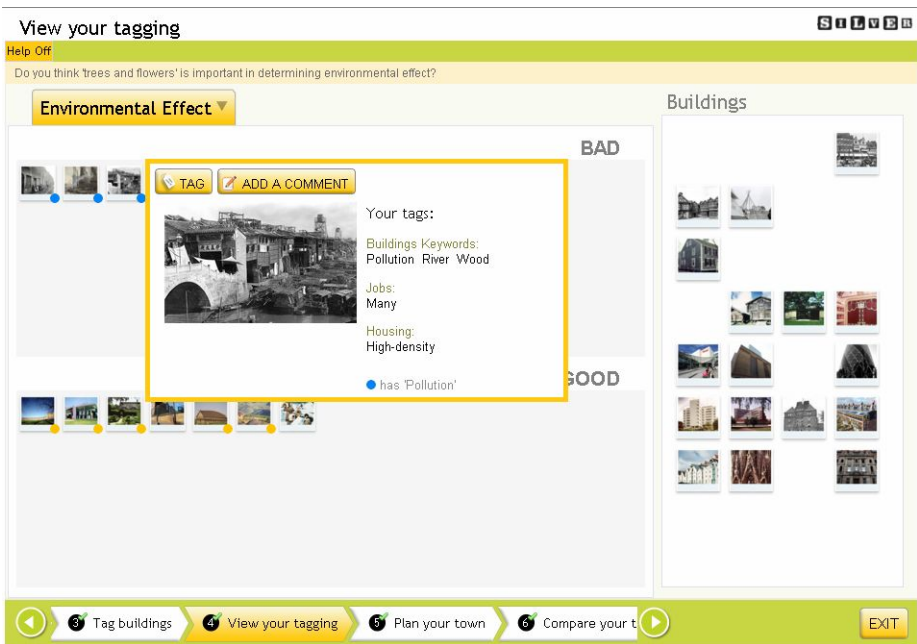
**Fig 1.** Tagging a building

about that portion of the image, or it may link to text, movie or audio, but related to just that part of the image rather than the picture as a whole. Based on the attributes they have applied to a building, the user organises the resources into categories which are related to the question they are answering: in this case categorising a resource as either good or bad for environmental, economic or social sustainability.

Figure 2 shows an example of the ‘view your tagging’ stage, in which several images have been categorised according to whether they are good or bad for environmental effect. This page represents a visual summary of the user’s input from the “tag buildings” task. Spatially, the thumbnail images are grouped according to their classification. Further visual information is provided via coloured markers attached to the thumbnails. Hovering over an image in this view displays the key for a dot. In this case, the blue dots represent resources that have been tagged as having “pollution”. The pattern that has emerged is that the user has been categorising images showing pollution as being “bad” for environmental effect. The information bar below the “Help” button provides additional prompts relating to the patterns – in figure 2, this information bar displays the question “do you think ‘trees and flowers’ is important in determining environmental effect?” While on this screen, users are encouraged to reflect on the patterns they have created between their tagging and their classifications and to consider any inconsistencies that have been highlighted via the coloured markers. For example, the user might notice that all but one of the resources in the

category ‘bad for environmental effect’ have a blue dot representing ‘pollution’. This could be due to incorrect tagging, incorrect classification, or to the resource being an exceptional case. The user can interact directly with resources on this screen to:

- change tags: clicking on the “tag” button in the resource box brings up a screen where the user can re-tag the resource.
- change classification: the user can drag and drop the resource to a different category.
- comment: the user can make a comment on the resource (e.g. to explain it’s classification) by clicking on the comment box in the resource window. The comment might be for the user alone, or else to share with other people in the group who are also sharing the resources.



Produced by Bridgeman Education, Lexara Ltd and The Knowledge Media Institute. Funded by the Department for Innovation, Universities and Skills (DIUS)

**Fig. 2.** Classifying environmental effect. A blue dot represents pollution (the yellow dot represents trees and flowers).

## 4 Generation and Presentation of Feedback

The data for the augmented visualisation is provided by the SILVER-RE. The ID3 algorithm [8] generates a decision tree, based on all the input data for the specified class. In the sustainability example, the class might be a single dimension, e.g.:

environmental effect = good  
environmental effect = bad

or a combination of dimensions, e.g.:

environmental+social effect = good+good  
 environmental+social effect = good+bad  
 environmental+social effect = bad+good  
 environmental+social effect = bad+bad

At each step of tree generation the ID3 algorithm selects an attribute on which to split, based on which will most likely predict the correct classification for the given data set. The decision at each node is only locally optimal, since the algorithm does not look ahead. Therefore, to improve the data provided by ID3, the SILVER-RE produces multiple trees by additionally selecting nodes where the entropy value at the node falls within a specified threshold. The resulting decision trees can be turned into a set of rules, in an If-THEN format, e.g.

IF pollution THEN environment = bad  
 IF grass and flowers and trees and not pollution THEN environment = good

Each rule describes a path from the root of the tree to a leaf node. The rules are pruned to remove duplicates and also non-informative-attributes, e.g. where all attribute values are expressed across otherwise identical rules or where the case can be covered by a more general rule.

Further post-processing is carried out on the rule-set to produce data that can be used to provide meaningful feedback to the user. The type of feedback given depends on the task context. This is specified via an XML request from Magic Studio (MS). Once processed by the SILVER-RE, the result is returned to MS within an XML template, whereupon it is used to augment the current visualisation.

To generate the feedback shown in figure 2, the SILVER-RE determines, from the ID3 generated rule-set for the current data, the top 'n' attributes that discriminate between the classes: in this case, environmental effect = good, environmental effect = bad. Each attribute returned can have two or more values. Attributes are visualised by colour, for example:

pollution=blue  
 trees and flowers = yellow  
 jobs=red

The values of an attribute is visualised by shape, for example:

blue circle = pollution=yes  
 blue square = pollution=no  
 red circle = jobs+none  
 red square = jobs+some  
 red triangle = jobs+many

In order to keep the visual summary concise and informative, the number of colours and shapes is limited. Some of this processing is done by Magic Studio, for example where an attribute has only a possible “yes” or “no” value, the ‘no’ values are not visualised. Other processing is done via the SILVER-RE which can be adjusted to return an appropriate number of discriminatory attributes. For some tasks, this might be a maximum of two attributes with three possible values each, but if all attributes are binary then three might be more appropriate.

The reasoning engine can offer further functionality not yet supported by the Magic Studio interface. This has been demonstrated through a separate tool for visualising SILVER tasks, which has been developed in Java. This works to the same XML specification as Magic Studio, but includes additional visualisation functionality, some of which is now described.



Fig. 3. Tag cloud

One additional function is the ability to produce a tag cloud (figure 3) that can be used to visualise the frequency of tags applied in different classes and to further highlight which of those tags are discriminatory attributes for the task, or which might be incidental (e.g. common attributes that appear in all classes). Attributes that discriminate are shown in bold.

The java tool can visualise the top ‘n’ attributes of multiple dimensions. This can be done by a table or, in certain cases (i.e. when there is a maximum of 3 dimensions, none of which have more than two possible values), by a Venn diagram (figure 4). Feedback is requested every time the user interacts and makes a change to either a tag or a classification. The returned result is visualised immediately. This means that if the top ‘n’ attributes change then the meaning of the coloured markers can change too.

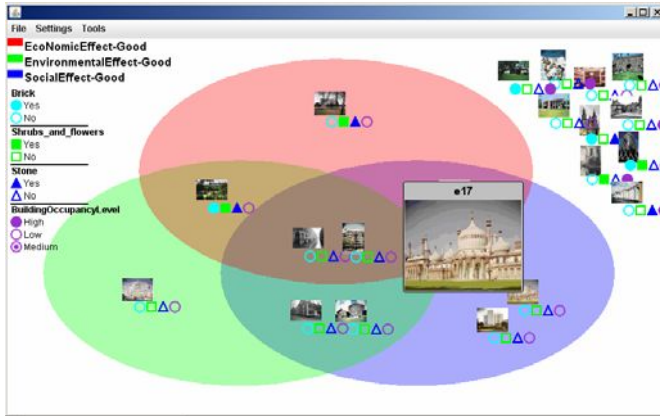


Fig. 4. Visualising 3 dimensions of sustainability on a Venn diagram

## 5 Classroom Assessment

Two classroom assessments have been carried out, with 13 students in one school and 27 students in another. The assessments were both carried out as part of the prototype design stage, the purpose being to understand how teachers would use the tool in the classroom and to watch students interacting directly with the tool and to understand and address any problems that they had with it. As such, the sessions were informal and the data gathered was, out of necessity, observational in nature since the intention was to allow teachers to use the tool within a lesson in a way that was natural to them, rather than asking them to comply with strict experimental procedures.

In each assessment, students interacted with the prototype, under the guidance of their teacher, within a lesson on sustainability. Students tagged several buildings and then viewed the feedback that had been created by the SILVER-RE. Talking to the students at this stage revealed that the majority clearly understood the patterns that they had created. Some students would try to create further explanations for their patterns, such as noticing that their top environmental attributes were all ‘green things’, or that their economic attributes were things that made money. Some students might then continue to reorganise their resources into different categories by dragging and dropping them to see how the pattern might change.

Also clear was the tendency of students and teachers to generate causal or dynamic models in explaining or elaborating the patterns found and presented in the visualizations. For example, ‘place for trade, business and tourism’ occurring in cases that are classified as economically sustainable would generate an explanation around how trade and tourism would increase jobs and the income of residents of the area, which would then be spent in local shops. This shift from classification to dynamic model descriptions is possible because each of these cases can essentially be thought of as inputs and outputs of a system and the explanation is essentially adding structure to the behavioural description of that system snapshot. This finding led us to the conclusion that we should include some form of associated system model in order to help

prompt interpretation of results in terms of a dynamic model and guide the selection of attributes within visualizations.

As we describe in the next section, we have developed an approach to qualitative modelling that can be used to specify models that support inquiry learning from sources. Qualitative modelling has been found to be applied as a pedagogical technique to support students in understanding and expressing relationships between system components (e.g. [9, 10]). Qualitative modelling has been found to be more accessible than, as well a bridge to, more quantitative methods of expressing systems and their behaviour [11].

## 6 Conceptual Model

In the sustainability task, we are using qualitative models to guide inquiry learning from sources. However, the task of interrogating sources can also be seen as a potential way of uncovering or discovering relationships within a qualitative model that the learner may have difficulty appreciating or articulating, but may be able to demonstrate through a knowledge building activity. Berry and Broadbent [12] showed how there can be a dissociation between task performance and related verbalizable knowledge. Techniques employed for knowledge acquisition [13] such as repertory grids also demonstrate this principle that systems relationships can sometimes only be demonstrated instead of, or as a precursor to their articulation. This suggests that this association between a model and inquiry learning from sources could have a dual benefit. The model could provide guidance for the inquiry learning task. Additionally, the inquiry learning task could act as a precursor to understanding aspects of the model. The next section describes our approach to specifying associated qualitative models, in the context of a task for learning about leadership styles. First, the task is described and then the conceptual model.

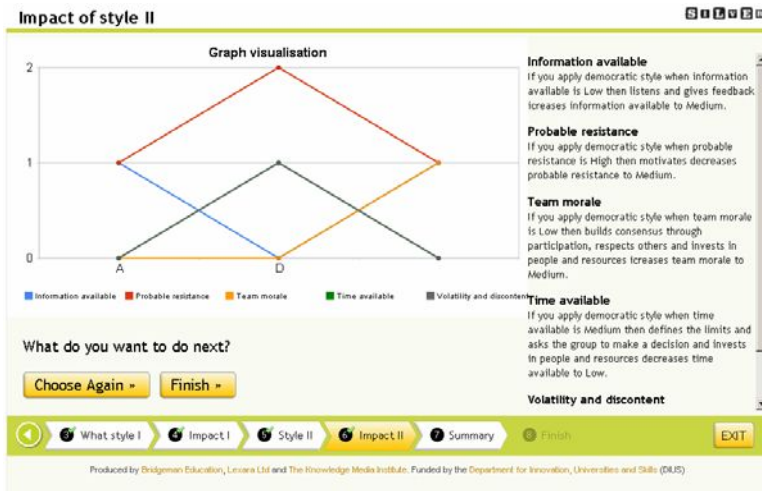


Fig. 5. Visualising the effect of choosing different leadership styles



In the leadership module, a user first learns about different leadership styles, such as autocratic, democratic or transformational. For each style, the user is provided with information about the behaviours exhibited by those styles and the sorts of situations in which it is or isn't appropriate to use it. For example, in an autocratic style, the leader 'decides and orders'. This behaviour is useful for making the most of the available time and so should be used if time is short, but it can cause unrest and should be avoided if staff are already upset and discontented. Currently, this part of the module is provided by standard text/pictures, although it is possible to support it through a SILVER type tagging and classification task. Once the user is familiar with the different styles and when they should be used, they are presented with complex scenarios that requires several styles to be applied to address different aspects of the task. In a complex scenario, the style applied in each step has some effect on the scenario parameters which may affect the user's choice in the next step. The aim for the user is to explore the trade-offs of each particular style, e.g. 'listening to staff and pooling ideas increases information available for problem-solving but takes time', and learn that there is rarely one style that is suitable to resolve any given situation. When participating in the scenario task, the user reads, or views a video showing the whole scenario and then decides which scenario attributes apply, from:

- Information available (to you)
- Probable resistance (to your actions)
- Team morale
- Relative time
- Volatility and Discontent

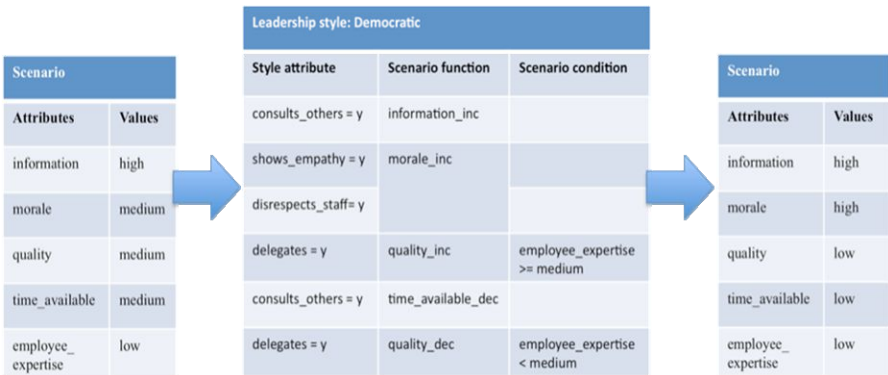
The effects are visualised on a graph so the user can see the scenario attributes going up and down as they select different styles. Each time the user chooses a style and visualizes the result, they have the option to redo the step, choose a new style to apply based on the current state, or to end the task, if they feel the outcome is the best they could achieve. The user can choose a maximum of 5 styles for any scenario. An example graph is shown in figure 5.

This task is supported by a generic conceptual model that has been instantiated with leadership task properties. The model describes the effect that certain leadership style attributes have on aspects of the scenario that the user is trying to influence, such as the available time, the morale of employees etc. A behaviour can cause an attribute to go either up or down, within the specified, but modifiable, range, e.g. "low, medium, high" or "1 to 5". Figure 6 shows the permissible transitions between values, either incrementing or decrementing.



**Fig. 6.** Representation of how scenario attributes can be incremented and decremented

Once a scenario attribute reaches a limit, there are two possible outcomes depending on what has been pre-specified when setting up the model. In figure 10 these are represented via the presence or absence of arrows at the minimum or maximum values. In the first case, the scenario attribute can reach a limit, such as “low” and then either stay low, or else recover, based on the effect of the next style applied (this situation is represented by the presence of an arrow at the endpoint). In the second case, the scenario can reach a limit beyond which it cannot recover and effectively the task is “failed” (represented by the absence of an arrow at the endpoint). For example, if a leadership style fails to maximize on the available time in each step, then time can run out to complete a task.



**Fig. 7.** A style has a set of attributes that has associated functions (middle). These transform on set of scenario attributes (left) into another (right).

Similarly, if a leadership style effects the volatility and discontent of your team, then it can reach a limit where the team walk out and refuse to work. As shown in figure 5, the effect of a style choice is explained, in terms of the underlying model, in the right-hand pane.

Figure 7 shows an example of how the model maps style behaviours to functions that increase or decrease scenario attributes.

## 7 Conclusions and Future Work

Principles of spatial hypertext have been used to visualise patterns in data in a concept learning task. The SILVER functionality can be used to:

- show how a set of resources have been tagged and classified by a user.
- use machine learning to identify patterns between tags and resources and identify tags which, in the user’s point of view, best discriminate between classes.
- highlight inconsistencies in the patterns, which could be due to incorrect classification, incorrect tagging, or the resource being a special example.
- allow the user to directly interact with the summary to see how changes to the data can change the pattern.

- visually compare the tagged resources of two different users and suggest which changes to either tagging or classification would have the biggest impact on aligning one users point of view with the other.
- use surrogates to refer to anything to which tagging and classifications can be applied, such as any media types, or different people within a group.

Much of this functionality has been evaluated within the framework of a task on sustainable development. The evaluations have suggested that students are able to interpret the visual summaries in the context of the task.

A conceptual model has been included in the next instantiation of the SILVER technology, to allow better explanations to the user. Future work on the conceptual model will involve allowing users to achieve the best result for specified scenario attribute by “running” the model and finding what the best choice of styles would be to achieve a selected result, e.g. to get the best result for ‘time’ or the best for ‘time and employee morale’. The leadership prototype will soon be tested online with the target audience, which is young professionals just starting their careers and undertaking some leadership training as part of their career development, or undergraduates doing business degrees.

## References

1. Burke, M.: *Organisation of Multimedia Resources: Principles and Practice of Information Retrieval*. Gower Publishing Company, Hampshire (1999)
2. Shipman, F., Hsieh, H., Airhart, R., Maloor, P., Moore, J.M.: *The Visual Knowledge Builder: A Second Generation Spatial Hypertext*. *ACM Hypertext*, 113–122 (2001)
3. Marshall, C., Shipman, F.: *Spatial hypertext and the practice of information triage*. In: *HYPERTEXT 1997: Proceedings of the eighth ACM Conference on Hypertext*, pp. 124–133. ACM Press, New York (1997)
4. Shipman, F., Moore, J.M., Maloor, P., Hsieh, H., Akkapeddi, R.: *Semantics Happen: Knowledge Building in Spatial Hypertext*. In: *Proceedings of Hypertext 2002*, pp. 25–34 (2002)
5. Scardamalia, M., Bereiter, C.: *Knowledge building*. *Encyclopedia of Education*, pp. 1370–1373 (2003)
6. Hsieh, H., Shipman, F.: *Supporting Visual Problem Solving in Spatial Hypertext*. *Journal of Digital Information* 10(3) (2009)
7. Buchanan, G., Blandford, A., Thimbleby, H., Jones, M.: *Integrating information seeking and structuring: exploring the role of spatial hypertext in a digital library*. In: *Proceedings of Hypertext 2004*, pp. 225–234 (2004)
8. Quinlan, J.R.: *Induction of decision trees*. *Machine Learning* 1(1), 81–106 (1986)
9. Bouwer, A., Bredeweg, B.: *VisiGarp: Graphical Representation of Qualitative Simulation Models*. *Artificial Intelligence in Education, Japan, Osaka* (2001)
10. Gupta, R., Wu, Y., Biswas, G.: *Teaching about Dynamic Processes A Teachable Agents Approach*. *Artificial Intelligence in Education, Amsterdam, The Netherlands* (2005)
11. Bredeweg, B., Forbus, K.: *Qualitative modelling in education*. *AI Magazine* 24(4), 35–44 (2003)
12. Berry, D.C., Broadbent, D.E.: *On the relationship between task performance and verbalizable knowledge*. *Quarterly J. of Experimental Psychology* 36A, 209–231 (1984)
13. Boose, J.H., Gaines, B.R.: *Knowledge Acquisition for Knowledge-Based Systems: Notes on the State-of-the-Art*. *Machine Learning* 4, 377–394 (1989)

# Building the Knowledge Base to Support the Automatic Animation Generation of Chinese Traditional Architecture

Gongjin Wei<sup>1</sup>, Weijing Bai<sup>1</sup>, Meifang Yin<sup>1</sup>, and Songmao Zhang<sup>2</sup>

<sup>1</sup> Beijing Municipal Key Laboratory of Multimedia and Intelligent Software Technology,  
Beijing University of Technology, Beijing 100022, China

{weigongjin, estherbaiweijing, meifangyin119}@gmail.com

<sup>2</sup> Academy of Mathematics and Systems Science, Chinese Academy of Sciences,  
Beijing 100190, China  
smzhang@math.ac.cn

**Abstract.** We present a practice of applying the Semantic Web technologies in the domain of Chinese traditional architecture. A knowledge base consisting of one ontology and four rule bases is built to support the automatic generation of animations that demonstrate the construction of various Chinese timber structures based on the user's input. Different Semantic Web formalisms are used, e.g., OWL DL, SWRL and Jess, to capture the domain knowledge, including the wooden components needed for a given building, construction sequence, and the 3D size and position of every piece of wood. Our experience in exploiting the current Semantic Web technologies in real-world application systems indicates their prominent advantages (such as the reasoning facilities and modeling tools) as well as the limitations (such as low efficiency).

**Keywords:** Knowledge representation and reasoning, the Semantic Web, ontology, rules, AI-based animation, Chinese traditional architecture.

## 1 Introduction

The Semantic Web technologies provide knowledge representation and reasoning mechanisms that enable computer systems and agents to understand the meaning of the web sources and further facilitate the communicate and interoperation among these systems [1]. The foundation of the Semantic Web is the ontology that represents a shared conceptualization of a given domain. Description logics-based ontologies have decidable reasoning support, and the OWL languages are the standard for specifying ontologies. Built upon the ontology on the Semantic Web are the more powerful representation formalisms including the rule languages. Over the past ten years, research on the Semantic Web has been largely developed with the presence of various formalisms, frameworks, algorithms and tools [2]. Application of these technologies in real-world software systems is essential for the success of the Semantic Web.

We present in this paper a practice of applying the Semantic Web technologies in the domain of Chinese traditional architecture. More precisely, we built a so-called

full life-cycle automatic animation generation system, where the user describes what kind of building he or she wants to see and the system accordingly generates an animation to demonstrate the construction of the building in real time. A knowledge base about the Chinese traditional architecture was modeled to support the whole automation from the user's input to the animation. We used ontology and rule languages to capture the domain knowledge so as to make the knowledge base re-usable and shareable on the Semantic Web.

The paper is organized as follows. Section 2 gives three diverse backgrounds that our work is related to, including the full life-cycle animation generation technology, the application of the Semantic Web in the architecture domain, and the Chinese traditional architecture. Section 3 presents in details the ontology and the rules to serve the different purposes in the automatic animation generation. Two implementations of the knowledge base are described in Section 4 as well as the result of the efficiency improvement. In section 5 we conclude the advantages and drawbacks of our work. Of note, no novel Semantic Web techniques are proposed in this paper. Rather, we present our experience in applying the Semantic Web techniques in a new domain, i.e., the automatic generation of animations for building Chinese traditional houses. The experience shows the feasibility as well as the limitations of the Semantic Web technologies in real-world software systems.

## **2 Background**

### **2.1 Full Life-Cycle Automatic Generation of Animation**

As an AI-based animation approach [3], Ruqian Lu proposed the full life-cycle computer-aided automatic animation generation technology in the late 80's [4]. His team implemented a system to demonstrate the technology in the mid 90's [4]. Starting from a story written in some limited Chinese natural language, a system called SWAN covered the complete automatic process consisting of natural language processing, story understanding, animation plot design, character design, action planning, camera control and finally, generation of a 3D animation.

Based on this technology, we built a full life-cycle automatic animation system for Chinese traditional architecture [5]. The modules in the automation include information extraction of the user's description of the architecture, reasoning of the components needed for the architecture and their construction sequence, computation of the size and position of the components in a 3D scene, generation of a qualitative representation of the movements of these components, generation of joints in a 3D scene, generation of a quantitative representation of the movements, and lastly generation of animation files ready for rendering. For details of our animation system please refer to [5]. In this paper we focus on the design and implementation of the Semantic Web knowledge base in this animation system.

## 2.2 Applying the Semantic Web Technologies in the Architecture Domain

Although the Semantic Web technologies have been used largely in many domains such as biomedicine [6], little work is presented about the architecture domain. We give a brief introduction to two related studies here. The first [7] uses one of the modular ontology formalisms,  $\mathcal{E}$ -connections [8], to specify different ontological layers of architectural design. The resulting quantitative layer, qualitative layer, and a conceptual layer represent the same building structures from different aspects.  $\mathcal{E}$ -connections links define the semantic connections among elements across these ontology layers. Using the reasoner of  $\mathcal{E}$ -connections can ensure the semantic consistency among these ontologies. The second work [9] constructed a Chinese traditional architectural element library for converting various types of point cloud CAD data into 3D modeling. Rules were developed to specify the constraints for shape, size, position, rotation and connection of the wooden elements in an architecture. We will compare our work with these two studies in the conclusions section.

## 2.3 Chinese Traditional Architecture

Chinese traditional architecture is unique in the world and the construction of the timber structure is the core of the building. Generally speaking, a timber structure consists of three primary levels, the frame network of columns, the Dougong<sup>1</sup> level, and the roof. The wooden pieces in the timber structure connect together by joints alone to form a stable structure. Ancient Chinese central governments actually released treatises [10] to regulate the ranks of architecture and the materials used. For example, a Wudian timber structure is of the highest rank and only royal buildings were permitted to adopt the Wudian style. The Taihe Temple in the Forbidden City is a typical complex Wudian building with a double-layered roof.

# 3 The Ontology and the Rule Bases

The knowledge base we built to support the automatic animation includes an ontology and four rule bases. The former conceptualizes the domain of the Chinese traditional timber structures, and the latter specify the components needed for building a given timber structure, construction sequence of these components, and the computation of their 3D size, position and rotation.

## 3.1 The Ontology of the Chinese Traditional Timber Structures

We built an OWL DL ontology to represent the entities of the Chinese traditional architecture and the relationships among these entities. More precisely, classes of the ontology include types of the timber structures, and types of pieces of wood that the

---

<sup>1</sup> Dougong is a “unique structural element of interlocking wooden brackets” as described in <http://en.wikipedia.org/wiki/Dougong>

timber structure is composed of. The subclass relationships between classes represent a classification of the entities in the ontology. The following are three of such child-parent relations.

$$\begin{aligned} DoubleLayeredWudian &\sqsubseteq Wudian \\ Wudian &\sqsubseteq TimberStructure \\ InnerColumn &\sqsubseteq Column \end{aligned}$$

They represent that a Wudian structure with double layers is a type of Wudian structure, which is a type of timber structure, and inner columns are a type of columns. In addition to subclass relationships, we also specified partitive relationships between classes of the components and classes of the timber structure. For example, the axiom

$$Wudian \sqsubseteq \exists \text{ hasPart. InnerColumn}$$

represents that every Wudian structure has at least one inner column as its component. Moreover, we defined associative relationships between wooden components. The following axiom specifies that for every purlin from left to right, there is at least one purlin from front to rear so that the two components are connected through a joint. The relationship *hasJointWith* is defined to be symmetric.

$$LeftToRightPurlin \sqsubseteq \exists \text{ hasJointWith. FrontToRearPurlin}$$

The description logic reasoner Pellet<sup>2</sup> was used to ensure the logical consistency of the ontology.

### 3.2 The Rule Base for Generating Individuals for Every Piece of Wood

To demonstrate the construction process we take into account every single piece of wood installed in the timber structure. In order to specify such wood, we dissect the timber frame structure in three directions, vertical, horizontal from left to right, and horizontal from front to rear. Further we represent one piece of wood through a 4-tuple, i.e., the type of wooden component the wood belongs to, the start and end layer it comes across vertically, left-to-right horizontally, and front-to-rear horizontally, respectively. The 4-tuple representations of pieces of wood are important for deciding their construction sequence and their 3D sizes, positions and rotations. Take for example a Wudian building with  $n$  left-to-right sections,  $m$  purlins on the roof and a veranda<sup>3</sup>. All its inner columns at the left-to-right horizontal direction are represented as:

$$(InnerColumn, (1,1), (j,j), (3,3)), (InnerColumn, (1,1), (j,j), (M,M))$$

where  $M$  is the number of purlins (i.e.,  $m$ ), and  $j$  traverses from  $(J'+1)$  to  $(J-J')$  where  $J'=I+(M+1)/2$  and  $J=M+N+2$ ,  $N$  is the number of sections (i.e.,  $n$ ).

We built a rule base to represent the knowledge about, for example, how many inner columns are needed in a Wudian building of  $n$  sections,  $m$  purlins and a veranda.

<sup>2</sup> <http://clarkparsia.com/pellet>

<sup>3</sup> A veranda is a roofed open porch that surrounds the exterior walls and is optional for the Wudian style building.

We use Jess [11] to do the rule reasoning, so the rule examples<sup>4</sup> in this paper are presented in the Jess format. Based on the data extracted from the user's description of the architecture, the knowledge base system infers what types of wooden components are needed and their respective numbers, and accordingly generates individuals for the corresponding classes in the ontology. The following rule generates the individuals of inner columns, where *bind* assigns values (e.g., *I*) to variables (e.g., *jzid* which is a variable for the ID of individuals).

```
(bind ?j (+ ?J1 1))
(bind ?jzid 1)
(while (<= ?j (- ?J ?J1)) do
  (assert (InnerColumn (ID ?jzid) (VerticalLayerFrom 1) (VerticalLayerTo 1)
    (leftToRightHorizontalFrom ?j) (leftToRightHorizontalTo ?j)
    (frontToRearHorizontalFrom 3) (frontToRearHorizontalTo 3)))
  (assert (InnerColumn (ID (+ ?jzid 1)) (VerticalLayerFrom 1) (VerticalLayerTo 1)
    (leftToRightHorizontalFrom ?j) (leftToRightHorizontalTo ?j)
    (frontToRearHorizontalFrom ?*M*) (frontToRearHorizontalTo ?*M*)))
  (bind ?j (+ ?j 1))
  (bind ?jzid (+ ?jzid 2))
)
```

Such rules infer that, e.g., for a Wudian building with five sections, nine purlins and a veranda, 26 types of wooden components are needed and a total of 461 pieces of wood are generated, including 16 individuals of the inner column class and 24 individuals of the outer column class.

### 3.3 The Rule Base for Computing the 3D Size

The animation generated demonstrates the construction of a building in a 3D scene where pieces of wood fall one-by-one in certain sequence into their final positions in the architecture. We studied the architectural treatises of the ancient Chinese central governments [10] and summarized rules to compute the size of every type of wooden component, i.e., the three size values at the vertical, left-to-right horizontal and front-to-rear horizontal direction. The shape of wooden components can be rectangular, cylindrical or irregular, and for the irregular we compute the size of its bounding rectangle.

**Computing the sizes in three directions.** We take for example a special kind of wooden components, called *Gua*-column, as shown in Figure 1. *Gua*-columns are an important supporting component sit between two layers of beams on the roof. The sizes of a *Gua*-column at three directions are closely related to the size of the beam underneath it, e.g., the front-to-rear horizontal width of a *Gua*-column is obtained from subtracting 0.4 *Doukou* from the left-to-right horizontal width of the underneath beam, where *Doukou* is one of the measure units used in ancient China for buildings. The following rule computes the sizes of the *Gua*-column.

---

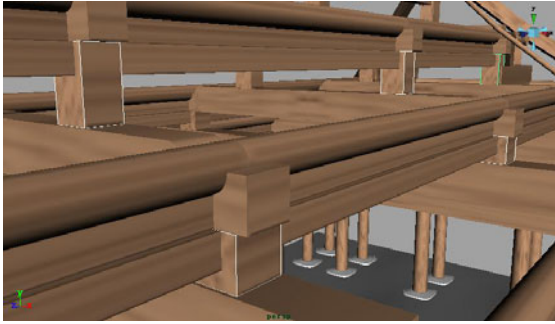
<sup>4</sup> Please note that the rule examples presented in this paper can not be fully understood since the complete rule bases are not given. Domain knowledge is also required. We use the rule examples to illustrate the complexity of the modeling.



```

(defrule size_GuaColumn
  ?g <- (GuaColumn (VerticalLayerFrom ?cf&:(neq ?cf (div (- ?*m* 1) 2)))
    (LeftToRightHorizontalFrom ?mf))
  (Beam (VerticalLayerFrom ?cf2&:(eq ?cf2 (- ?cf 1)))(LeftToRightHorizontalFrom ?mf)
    (VerticalSize ?lcz)(LeftToRightHorizontalSize ?lmk))
  =>
    (bind ?mk (- ?lmk 0.8))
    (bind ?js (- ?lmk 0.4))
    (bind ?cz (nth$ ?cf ?CZ))
    (bind ?cz (- ?cz ?lcz))
    (modify ?g (VerticalSize ?cz)(LeftToRightHorizontalSize ?mk)(FrontToRearHorizontalSize ?js))
)

```



**Fig. 1.** Gua-columns in white lines in an animation for the construction of a Wudian structure generated by our system

**Computing additional lengths for the joint connection.** For the purpose of joint connection, additional length is needed for some wooden components. Take the jointing purlins described in section 3.1 for example. In order for the two purlins to cross-connect, their lengths should include what is needed for their joint. The following rule is for computing such purlins at the left-to-right direction, based on other components including some of the stringers at the left-to-right direction, and the left-to-right size of *YouQiang*.

```

(defrule size_LeftToRightPurlin2
  ?f <- (or (Purlin (LeftToRightFrom ?mf&?*!ForPurlin*))
    (Purlin (LeftToRightTo ?mt&?*sForPurlin*)))
  (Data (Name Purlin)(Vertical ?cz)(FrontToRear ?js))
  (Data (Name YouQiang)(LeftToRight ?yq))
  =>
    (bind ?mf (fact-slot-value ?f LeftToRightFrom))
    (bind ?mt (fact-slot-value ?f LeftToRightTo))
    (bind ?mk (LeftToRightSpan ?mf ?mt))
    (bind ?add (* (+ (/ ?cz 2) ?yq) 2))
    (bind ?mk (+ ?mk ?add))
    (modify ?f (VerticalSize ?cz)(LeftToRightSize ?mk)(FrontToRearSize ?js))
)

```

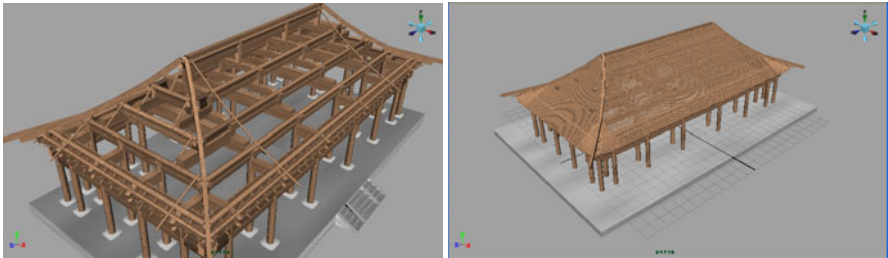
For the Wudian type of timber structure, we have totally summarized 40 size computation rules for 26 types of wooden components. These resulted in the computation of, for example, 1,059 pieces for a building with eleven sections and thirteen purlins, as shown in the middle column of Table 1.

**Table 1.** Number of pieces of wood computed by size and position rules

Wudian structures	Number of pieces of wood computed by the size rule base	Number of pieces of wood computed by the position rule base
3 sections, 7 purlins	285	1,115
5 sections, 7 purlins	369	1,493
7 sections, 7 purlins	453	1,027
5 sections, 9 purlins	461	1,940
7 sections, 9 purlins	563	2,443
7 sections, 11 purlins	673	3,213
9 sections, 11 purlins	793	4,073
11 sections, 13 purlins	1,059	5,539

### 3.4 The Rule Base for Computing the 3D Position

Once the size is obtained, the position of every single piece of wood can be computed. We built a rule base to compute the three axis values of the center of the wood in a 3D scene as well as the three rotation values. We take rafters (as in Figure 2) for example in the following to illustrate one of the most complex position computations.



**Fig. 2.** Rafters in an animation for the construction of a Wudian structure generated by our system; the figure on the left shows some of the key rafters, and the one on the right shows the complete rafters

**Computing rafters at the dissected points.** Rafters are installed side-by-side in parallel on the top of the roof, and only those positioned at the dissected points can be represented by our 4-tuple specification, called key rafters.

**Copying rafters.** Rafters between two key rafters, rather than in the first step for component generation, are not generated until now so that they can have the same size as the key rafters, and their number and positions are computed based on the size and position of the key rafters.

**Computing rafters at the corners.** Rafters at the four corners of a building have to be displayed to fit into an area of triangle. These corner rafters differ not only in position but also in size. They can not be generated or computed until the size of the area of triangle is known. We give a part of a rule for computing the corner rafters as follows, where *assert* is for generating rafters. The rule computes the area where the corner rafters are put, the rafter number, the distance between every two rafters, the progressively decreasing length of the rafters, and the three position values of the rafters.

```

.....
(bind ?pos1 ?leftToRightPos1)
  (bind ?pos2 (nth$ (- ?m1 1) ?*leftToRightPos*))
  (bind ?a ?rafterSize)
  (bind ?A (* (call Math Pi) (/ ?rM 180)))
  (bind ?allDistance (call Math abs (- ?pos1 ?pos2)))
  (bind ?radiusC (/ ?rafterSize 2))
  (bind ?num (call Math floor (/ ?allDistance (* ?radiusC 3))))
  (bind ?oneDistance (/ ?allDistance ?num))
  (for (bind ?i 1) (<= ?i ?num) (++ ?i)
    (bind ?reduce (/ (* ?a ?i) ?num))
    (bind ?newLength (- ?sC ?reduce))
    (bind ?newLeftToRightPos (+ ?leftToRightPos2 (* ?i ?oneDistance)))
    (bind ?newFrontToRearPos (+ ?frontToRearPos (* ?reduce 0.5 (call Math sin ?A))))
    (bind ?newVerticalPos (- ?verticalPos (call Math abs (* ?reduce 0.5 (call Math cos ?A))))))
  (assert (CopyRafter (leftToRightFrom ?m2)(LeftToRightTo ?m2)(FrontToRearFrom ?j1)
    (frontToRearTo ?j2)(VerticalLayerFrom ?cz1)(VerticalLayerFrom ?cz2)
    (LeftToRightSize ?sM)(FrontToRearSize ?sJ)(VerticalSize ?newLength)
    (LeftToRightPos ?newLeftToRightPos)
    (FrontToRearPos ?newFrontToRearPos)
    (VerticalPos ?newVerticalPos)(LeftToRightRotate ?rM)))
)

```

**Computing the outermost rafters.** The outermost rafters of a building are displayed so that a part of it hang in the air without support. The position of the center of the outermost rafter has to be adjusted so as to take the additional part into account. The following rule for the outermost rafters computes the rotation angles as well as the values similar to those mentioned above for corner rafters.

```

(bind ?CornerLength (+ ?DouGongLength 14))
(bind ?num (call Math floor (/ ? CornerLength (* ? rafterRadius 3))))
(bind ?oneDistance (/ ?allDistance ?num))
(bind ?xieBian (call Math pos(+ (* ?rafterLength ?rafterLength) (* ?cornerLength ?cornerLength))))
(for (bind ?i 1) (<= ?i ?num) (++ ?i)
  (bind ?reduce (/ (* (* ?xieBian (/ 1 3)) ?i) ?num))
  (bind ?newLength (- ?rafterLength ?reduce))
  (bind ?newLeftToRightPos (+ ?leftToRightPos2 (* ?i ?oneDistance)))
  (bind ?newFrontToRearPos (+ ?frontToRearPos (* ?reduce 0.5 (call Math sin ? rafterRotate))))
  (bind ? newVerticalPos (- ?verticalPos (call Math abs (* ?reduce 0.5 (call Math cos ? rafterRotate))))))
(bind ?angle (call Math (/ ?rafterLength ?cornerLength)))
(bind ?c1 (* ?cornerLength (- 1 (/ ?i ?num))))
(bind ?c2 (+ (* xieBian (/ 1 3) (- 1 (/ ?i ?num)))(* ?xieBian (/ 2 3))))
(bind ?newLength2

```

```

(call Math pow((-+ (*?c1 ?c1)(* ?c2 ?c2))(* 2 ?c1 ?c2 (call Math cos ?angle)))0.5)))
(bind ?rotate2
  (call Math acos(- (* ?c2 ?c2) (* ?c1 ?c1) (/ (* newLength2 newLength2) (* 2 newLength2
?c1))))))
(bind ?newLeftToRightPos2 (- ?newLeftToRightPos (* (/ 1 2) ?newLength2 (call Math cos rotate2))))
(assert (CopyRafter (leftToRightFrom ?m2)(LeftToRightTo ?m2)(FrontToRearFrom ?j1)
  (frontToRearTo ?j2)(VerticalLayerFrom ?cz1)(VerticalLayerFrom ?cz2)
  (LeftToRightSize ?sM)(FrontToRearSize ?sJ)(VerticalSize ?newLength2)
  (LeftToRightPos ?newLeftToRightPos2)
  (FrontToRearPos ?newFrontToRearPos)
  (VerticalPos ?newVerticalPos)(LeftToRightRotate ?rM)(FrontToRearRotate ?rotate2)))
)

```

For the Wudian type of timber structures, we have totally constructed 57 rules for position and rotation computation, and the number of components obtained is shown in the right column in Table 1. Take a Wudian building with eleven sections and thirteen purlins for example. The position rule base can infer 5,539 pieces of wood, where 4,480 of them are rafters whose size was actually computed by the position rule base rather than by the size rule base.

### 3.5 The Rule Base for Inferring the Construction Sequence

Constructing Chinese traditional architecture has to follow a certain sequence so that every two pieces of wood that meet together can be jointed correctly. We build a rule base to specify the construction sequence of components, for example, inner columns must be installed before the outer columns. The following rule represents that for the top level of the building, first install the *TaiPing*-beams, and then the *LeiGong*-columns, followed by the *Heng*-cross beams, and these three types of components have to satisfy the specified conditions, respectively.

```

(defquery topLevelTheFirstDissectSequence
  (declare (variables ?x)
    ?f1 <- (TaiPingBeam (VerticalLayerFrom ?*M5*) (VerticalLayerTo ?*M5*)
      (LeftToRightFrom ?a&:(eq ?a ?x)) (LeftToRightTo ?a)
      (FrontToRearFrom ?*I1*) (FrontToRearTo ?*I3*))
    ?f2 <- (LeiGongColumn (VerticalLayerFrom ?*M5*) (VerticalLayerTo ?*M5*)
      (LeftToRightFrom ?a) (LeftToRightTo ?a)
      (FrontToRearFrom ?*I2*) (FrontToRearTo ?*I2*))
    ?f3 <- (HengCrossBeam (VerticalLayerFrom ?*M5*) (VerticalLayerTo ?*M5*)
      (LeftToRightFrom ?a) (LeftToRightTo ?b&:(eq ?b (+ ?a 1)))
      (FrontToRearFrom ?*I2*) (FrontToRearTo ?*I2*))
  )
)

```

## 4 Implementations

Two implementations have been conducted for the knowledge base in our automatic animation system. In the first implementation, we used OWL DL (version 1.0) [12] to specify the ontology and the rule language SWRL [13] to represent the four rule

bases. More precisely, we modeled our knowledge base in the Protégé<sup>5</sup> environment where an OWL plugin and a SWRL Jess Tab are provided. The SWRL Rule Engine Bridge<sup>6</sup> enables the uniform reasoning based on the ontology and the rules. While using the advanced Semantic Web technologies, this implementation has the drawback of low efficiency. The middle column in Table 2 lists the time for inferring the construction sequence of various kinds of Wudian timber structures in Protégé<sup>7</sup>. Two to nearly eight minutes were needed for the sequence inference, which is only a small step in the whole automatic animation system. The low reasoning efficiency comes from that many tools were involved and the time was used for the frequent uploading of tools and data transformation.

**Table 2.** Efficiency comparison of the two implementations of the knowledge base

Wudian structures	Time for inferring construction sequence in Protégé (in minutes)	Time for inferring construction sequence in Jess (in minutes), and improvements
3 sections, 5 purlins	2.07	0.02 98.97%
5 sections, 7 purlins	2.52	0.02 99.18%
7 sections, 9 purlins	3.31	0.02 99.37%
9 sections, 11 purlins	4.89	0.02 99.52%
11 sections, 13 purlins	7.40	0.02 99.67%

This led to our second implementation where Jess was used to represent the whole knowledge base. Jess is a Java implementation of the CLIPS<sup>8</sup> production rule language for building expert systems. Jess uses the Rete algorithm for efficient reasoning which sacrifices memory for increased speed<sup>9</sup>. We converted the OWL DL ontology and the SWRL rules into the Jess rules. Classes and their subclass relationships in the architectural ontology are represented in Jess as exemplified as follows.

```
(deftemplate TimberStructure ... ...)
(deftemplate Wudian extends TimberStructure ... ...)
(deftemplate DoubleLayeredWudian extends Wudian ... ...)
```

The right column in Table 2 lists the time used for reasoning in the Jess-based implementation, which was saved up to 99%.

## 5 Conclusions

We have presented a knowledge base system consisting of one ontology and several rule bases. For the user's description of the structure of a building, the knowledge base system can infer what wooden components are needed and the number of pieces

<sup>5</sup> <http://protege.stanford.edu/>, version 3.4.4

<sup>6</sup> <http://protege.cim3.net/cgi-bin/wiki.pl?SWRLRuleEngineBridgeFAQ>

<sup>7</sup> The running environment is a PC with CPU 2.8GHz and memory 1.5G.

<sup>8</sup> <http://en.wikipedia.org/wiki/CLIPS>

<sup>9</sup> [http://en.wikipedia.org/wiki/Rete\\_algorithm](http://en.wikipedia.org/wiki/Rete_algorithm)

of wood for every type of components. Furthermore, the system computes the size and position of these pieces of wood in a 3D scene. The construction sequence can also be inferred to ensure the correct joinery of wooden components. From this step on, the animation modules can generate the qualitative and quantitative data about the movements of these pieces of wood in a 3D scene. Our animation system currently covers more than 180 types of Chinese traditional architecture.

Compared with other AI-based animation systems and tools, our technology has the feature of scalability. To expand the system to generate animations for another type of architecture, one only needs to summarize the rules for sequence reasoning and the rules of size and position computation, while all the other modules in the system can be used directly without updates.

Compared with the two studies aforementioned in section 2.2, what we have in common is the knowledge modeling of the architecture domain. The rules we summarized are for standard buildings based on the architectural treatises released by the ancient Chinese central governments. For real-world Chinese buildings, though, we have to manually transform the measuring data of the components into our representation. Methods for automatically extracting components from the point cloud data of a structure [9] could be complementary to our technology. Methodologically, the modular ontologies structure proposed in [7] could fulfill the similar functions as our knowledge base system does. However, the preliminary reasoning facilities for modular ontology languages make them currently infeasible in supporting real-world, large-scale knowledge bases.

We have implemented two versions of the knowledge base. The Protégé-based implementation takes full advantage of the friendly, flexible and integrated modeling environment, albeit it is inefficient in reasoning. On the other hand, the Jess-based implementation of the knowledge base is very fast in reasoning and computation, although modeling could be tedious and error-prone. Moreover, mixing ontological knowledge and rules together would impede the incremental development and scalability of the knowledge base. An automatic conversion from the OWL DL ontology and the SWRL rules into the Jess representation, as in [14], could enable us to keep the prominent features of the both sides. Moreover, incorporating our knowledge base under some upper-level ontologies or architectural standards would greatly facilitate its reuse and sharing.

**Acknowledgments.** This project is supported by the National Key Technology R&D Program of China, Knowledge Innovation Program of the Chinese Academy of Sciences, MADIS of the Chinese Academy of Sciences, and the Key Laboratory of Multimedia and Intelligent Software Technology at Beijing University of Technology.

We sincerely thank Prof. Chunnian Liu from Beijing University of Technology, and all of the graduate students participated in the design and development of the automatic animation system, including Kai Sun, Weifeng Wang, Liang Kong, Jia Sun, Lihuan Gong, Yan Wang, Jia Feng, Shebiao Liu, Tianzhu Liang, Bo Gu, Rongrong Sun, and Shan Zhu.

Lastly, we are very grateful to the anonymous reviewers for the helpful comments.

## References

1. Berners-Lee, T., Hendler, J., Lassila, O.: The Semantic Web. *Scientific American* 284(5), 34–43 (2001)
2. Staab, S., Studer, R.: *Handbook on Ontologies*. Springer, Heidelberg (2009)
3. Johansson, R., Berglund, A., Danielsson, M., Nugues, P.: Automatic Text-to-Scene Conversion in the Traffic Accident Domain. In: *Proceedings of 2005 International Joint Conferences on Artificial Intelligence (IJCAI 2005)*, pp. 1073–1078 (2005)
4. Lu, R., Zhang, S.: Automatic Generation of Computer Animation. LNCS (LNAI), vol. 2160. Springer, Heidelberg (2002)
5. Zhang, S., Sun, K.: Full Life-Cycle Automatic Animation Generation of Chinese Traditional Architectures. In: *2nd Asian Conference on Simulation and AI in Computer Games (GAMEON-ASIA 2010)*, pp. 63–68. EUROSIS-ETI Publication, Belgium (2010)
6. Bodenreider, O.: Biomedical Ontologies in Action. *Applied Ontology* 4(1), 1–4 (2009)
7. Hois, J., Bhatt, M., Kutz, O.: Modular Ontologies for Architectural Design. In: Ferrario, R., Oltramari, A. (eds.) *Formal Ontologies Meet Industry*. IOS Press, Amsterdam (2009)
8. Stuckenschmidt, H., Parent, C., Spaccapietra, S.: *Modular Ontologies: Concepts, Theories and Techniques for Knowledge Modularization*. LNCS, vol. 5445. Springer, Heidelberg (2009)
9. Zhu, L., Shi, R., Zhou, K.: Rule-based 3D Modeling for Chinese Traditional Architecture. In: *Proceedings of the 2nd ISPRS International Workshop on 3D Virtual Reconstruction and Visualization of Complex Architectures* (2007)
10. Guo, Q., Fashi, Y.: Twelfth-Century Chinese Building Manual. *Architectural History: Journal of the Society of Architectural Historians of Great Britain* 41, 1–13 (1998)
11. Friedman-Hill, E.: *Jess: the Rule Engine for the Java Platform*. Jess Manual (2008), <http://www.jessrules.com/jess/docs/71/>
12. McGuinness, D., van Harmelen, F.: *Web Ontology Language Overview*. W3C Recommendation (2004), <http://www.w3.org/TR/owl-features/>
13. Horrocks, I., Pater-Schneider, P.F., Boley, H., Grosz, B., Dean, M.: *SWRL: A Semantic Web Rule Language Combining OWL and RuleML*. W3C Member Submission (2004), <http://www.w3.org/Submission/SWRL/>
14. Mei, J., Bontas, E.P., Lin, Z.: OWL2Jess: A Transformational Implementation of the OWL Semantics. In: Chen, G., et al. (eds.) *ISPA-WS 2005*. LNCS, vol. 3759, pp. 599–608. Springer, Heidelberg (2005)

# Discovery of Relation Axioms from the Web

Luis Del Vasto Terrientes, Antonio Moreno, and David Sánchez

Universitat Rovira i Virgili. Departament d'Enginyeria Informàtica i Matemàtiques  
Intelligent Technologies for Advanced Knowledge Acquisition (ITAKA)  
Av Països Catalans, 26. 43007 Tarragona, Catalonia (Spain)  
{luismiguel.delvasto, antonio.moreno, david.sanchez}@urv.cat

**Abstract.** Given the proven usefulness of ontologies in many areas, the representation of logical axioms associated to ontological concepts and relations has become an important task in order to create an expressive representation of domain knowledge. Manual inclusion of logical axioms into an ontology can be a harsh, time consuming task. As a result, very few ontologies include axioms in their formal definition. From the ontology learning point of view, axiom learning is one of the less tackled and unexplored problems. In this paper we introduce a preliminary methodology to learn axioms associated to ontological relationships in an automatic and unsupervised way using the Web as corpus.

**Keywords:** Ontologies, Axioms, Knowledge Discovery, Semantic Web.

## 1 Introduction

An *ontology* represents the conceptual model underlying a certain domain, describing it in a declarative fashion and thus cleanly separating it from procedural aspects [2]. Ontologies play a key role in the Semantic Web [3] and other areas like agent communication, intelligent information integration and knowledge-based systems. Research in this area has increased considerably during the last years, bringing the development of modern ontological languages such as OWL 2, which provides a powerful formalism for knowledge representation and reasoning.

Ontologies are typically built by hand, requiring an arduous and time-consuming effort both from knowledge engineers and domain experts. As a result of the bottleneck introduced by knowledge acquisition approaches, manually composed ontologies are, in many cases, flat and lightweight, with very limited expressivity and mainly considering taxonomical knowledge and simple relations [4][5]. So, available ontologies very rarely exploit the expressiveness of languages like OWL 2 in defining logical constraints (axioms) associated to relations.

*Ontology Learning* methods, that extract automatically ontological elements from available corpora and build ontologies from them, mostly focus on taxonomic relationships [5] and, more rarely, on non-taxonomic knowledge [14]. Relation axioms involving logical connectives, role restrictions and other semantic features that can provide expressivity to ontologies remain largely unexplored [4].

This paper presents an automatic axiom learning algorithm which, starting from a set of non-taxonomic relations (denoted  $xRy$  or  $R(x,y)$ , in which  $x$  and  $y$  represent



concepts or individuals and  $R$  the relation/property) defined in an input ontology, explores and verifies which axioms those relations fulfill. Specifically, the axioms studied in our approach are *symmetry*, *reflexivity*, *functionality*, *transitivity* and *inverse*. Our approach relies on linguistic techniques and statistical analyses to unsupervisedly and automatically extract and evaluate semantic evidences found in Web resources [1], which support the definition of ontological axioms.

The paper is structured as follows. Section 2 presents the related works dealing with automatic ontology axiomatization. Section 3 describes our approach in detail, explaining the different stages of the learning process and the techniques used. Section 4 explains the criteria and measures considered to evaluate the algorithm and provides some preliminary results. Finally, the last section presents conclusions and some lines of future research.

## 2 Related Work

Several natural language processes have been applied to acquire OWL axioms from lexical resources in tools such as LEXO (that stands for Learning Expressive Ontologies), RELEXO (Relational Exploration) and HASTI, among others.

The implementation of LEXO basically relies on KAON2, an ontology management infrastructure for OWL, and the Minipar dependency parser [8]. Given a natural language definition of a class, LEXO starts by analyzing the syntactic structure of the input sentence. The resulting dependency tree is then transformed into a set of OWL axioms (concept inclusion, transitivity, role inclusion, role assertions, concept assertions and individual equalities) by means of manually engineered transformation rules [4]. The axiomatization is a semi-automatic cyclic process [8], in which each iteration contains an ontology evaluation step to make sure that the axiom candidates are correct. The main problem is the semantic inconsistencies caused by the analysis of natural language expressions. To tackle this problem the relational exploration [15] method was used in the development of the RELEXO application. This technique is an adaptation of attribute exploration from Formal Concept Analysis (FCA) to description logics, and is a good means to overcome the lack of completeness and precision in learned ontologies. It is natural to apply it to further specify the knowledge beyond the information extractable from the corpus, which makes relational exploration a perfect complement to automatic approaches for ontology generation based on lexical resources [7]. Another automated attempt was performed in [11], in which lexicosyntactic patterns are used to extract individuals and subclasses, and the OntoClean methodology is used to obtain evidence of strict disjointness between two classes. Pattern-extracted classes are analyzed and, if between two classes there aren't any common individuals or subclasses that taxonomically overlap, they are taken to be disjoint. In the OntoClean methodology, a pattern-based approach was used to analyze text taken from the Web specifically to verify that two classes have incompatible unity or identity criteria. Other patterns based in enumerations, defined in [9] for detecting disjointness, were not precise enough in comparison with average human disjointness detection.

HASTI is a system for automatic ontology construction, based on Persian (Farsi) text understanding. It uses an initial kernel, which has the essential metaknowledge (primitive concepts and operators) to build an ontology including axioms [6]. HASTI

implements a rule base which contains linguistic rules, inference rules and heuristics to extract lexical elements from texts. Specifically, it contains Semantic Templates to extract taxonomic relations, non-taxonomic relations and axioms. It extracts axioms (disjointness, transitivity and functionality) from conditional and quantified sentences, using axiom templates written in Knowledge Interchange Format (KIF). Following the lexico-syntactic approach, AuCONTRAIRE is a Contradiction Detection (CD) system which relies on the TextRunner system, an open Information Extraction system that uses the Web as corpus to obtain relations between entities [10]. In order to detect functional relations, it checks if a domain and its relation map to a unique variable using probabilistic assessment based on aggregated textual evidence.

The lack of automatic and unsupervised approaches to axiomatize ontologies motivated us to aim our efforts into this area. Some of these works include natural language patterns in order to determine axioms. In LEXO, RELEXO and the disjointness detection system there are rules that transform a set of patterns, determined by a particular natural language syntax, into axioms. For example, the Disjunction Rule is defined as a pattern “NP<sub>0</sub> OR NP<sub>1</sub>”, in which NP represents a noun phrase. If the text “Mazda or Toyota” is found, both brand concepts are set as disjoint. In our approach, the text patterns will not be only used to detect natural language structures from a text, but also to create Web search queries to extend the knowledge by using the biggest available repository. In this sense, on the contrary to these related works which are applied in a narrow context (i.e. a predefined document or even a sentence), our approach uses the Web as a massive learning corpus, enabling the discovery of relation axioms in a domain-independent fashion. On the other hand, HASTI uses structures of conditional sentences and compound sentences with a quantifier phrase to detect axioms, which may be a barrier to obtain complete information about the behaviour of the concepts and roles. Our proposed approach checks for evidences based only on the text pattern to analyse, avoiding the study of the sentence structure, taking as a matter of fact that logical axioms are not only stated on this type of sentences. AuCONTRAIRE uses text patterns and the Web as corpus, but its approach is based on detecting only contradictions by detecting functional relations. Unlike approaches such as LEXO and RELEXO, our approach does not require manual stages which, due to the Web size, would compromise its scalability.

### 3 Object Property Axioms Learning Methodology

*Object properties* define relations between pairs of ontological concepts. In contrast to hierarchical relations, an object property represents a non-taxonomic relation, which is typically expressed by a verb that relates a pair of concepts [12]. The definition of *object property axioms* allows the meaning of properties to be enriched and, therefore, the possibility to apply inference on them.

The pseudocode of the algorithm we propose is presented below, in which a domain ontology is received as input and a set of axioms associated to ontological relationships defined via object properties is obtained as output. First, the non-taxonomic relations from this ontology are read. For each relation, each potential axiom is analysed, generating a Web search query depending on a pattern create to analyse the axiom. The documents retrieved from the Web search query are analysed in order to find a match with the proposed pattern. Each match provides us with a term, which is

a candidate to satisfy a given relationship (e.g. a candidate to be an inverse relation). Every candidate is evaluated, using Web-scale statistics, to assess if it provides enough evidence to guarantee the fulfilment of the axiom. All these steps are explained in detail in the following subsections. During the explanation of the method, we refer to algorithm lines as *<line>*.

```

1. AxiomLearning (Domain Ontology DO)
2. { AXS_OP /* Results variable */
3.   AXIOMS /* Constant holding the set of axioms of interest */
4.   THRESHOLD /* Constant holding the selection threshold */
5.   RELS ← get_nontaxonomic_relations(DO)
6.   /* for each relation, analyze each object property axiom */
7.   for all RELSi ∈ RELS
8.     { for all AXIOMSj ∈ AXIOMS
9.       { QUERY ← generate_query(RELSi, AXIOMSj)
10.      WEBPAGES ← retrieve_WebPages(QUERY)
11.      for all WEBPAGESx ∈ WEBPAGES
12.        { /* for each web page check for matchings */
13.        TEXT ← extract_matchings(QUERY, WEBPAGESx)
14.        TEXT ← remove_stopwords(TEXT)
15.        KEY ← stem_word(TEXT)
16.        exists ← check_existence(AXS_OP, RELSi, AXIOMSj, KEY)
17.        if (!exists)
18.          { /* calculate web scale statistics */
19.          SCP = calculate_SCP(AXIOMSj, RELSi, TEXT)
20.          /* add info. to result if SCP > THRESHOLD */
21.          if (SCP > THRESHOLD)
22.            AXS_OP ← AXS_OP + (RELSi ∪ AXIOMSj ∪ KEY ∪ TEXT ∪ SCP)
23.          }
24.        }
25.      }
26.    }
27. Return AXS_OP
28. }
```

### 3.1 Pattern Construction

Lexico-syntactic patterns, as tools to discover and extract ontological entities, have been used in order to learn ontologies from unstructured text. Our approach bases the axiom learning in the exploitation of specially designed patterns aimed to retrieve (by means of querying a Web search engine) and extract (by means of a linguistic analysis) semantic evidences that support the fulfilment or not of each one of the studied Object Property Axioms. The patterns proposed below are domain-independent and can be adjusted to queries in many Web search engines. In base of these patterns, we have designed queries for the Yahoo! and Bing Web search engines <9>.

To describe the proposed patterns we use the nomenclature *Relation(Subject, Object)* in which *Relation* represents a relation and *Subject/Object* represent individuals. As an example, in a non-taxonomic relation defined as Borders(Spain, France), Spain is the *Subject*, France is the *Object* and Borders represents the *Relation*. In the following paragraphs the proposed patterns and their use to check some properties are described.

**Relation (Subject, ?).** This relation looks for possible *OBJECTS* for an individual *SUBJECT* and a *RELATION*.

*Functional Object Property & Inverse Functional Object Property:* Given a relation  $xRy$ , it is possible to detect if the subject  $x$  can be related to another object different from  $y$  by the relation  $R$  to check the *Functionality* of the Object Property  $xRy$  and the *Inverse Functionality* of the Object Property  $xR'y$  ( $R$  being the inverse of  $R$ ). In the query language we denote this pattern as [ $x R$ ] - [ $x R y$ ], in which [ $x R$ ] is used to extract new objects while [ $x R y$ ] avoids the extraction of the initial object  $y$  for the relation.

*Transitive Object Property:* Given a relation  $xRy$ , to detect the *Transitivity* of the property for the subject  $x$ , we need first to detect if the object  $y$  can be related to another individual  $z$  by relation  $R$ . In this way we can detect a relation  $yRz$  in which the relation  $R$  is the same for both  $xRy$  and  $yRz$ . In query language we denote this pattern as [ $y R$ ] - [ $y R x$ ]. [ $y R$ ] may help us to detect if  $y$  can be related to an object  $z$  by relation  $R$ , while [ $y R x$ ] avoids the extraction of the original subject  $x$  in  $xRy$  as object.

**Relation (Subject, Object).** We use this pattern to check if a particular relation  $R(x,y)$  holds.

*Symmetric Object Property:* Given a relation  $xRy$ , by switching the positions of the subject  $x$  and the object  $y$  we can check if the relation  $R$  can be applied from  $y$  to  $x$  as well, denoted in a web search query as [ $y R x$ ].

*Reflexive Object Property:* Given a relation  $xRy$ , to check its reflexivity we used as object the reflexive pronoun 'itself', which is used to refer to things or animals. In query language we denote this pattern as [ $x R$  itself].

*Transitive Object Property:* As said in the previous pattern Relation (Subject, ?), we can detect  $yRz$  from an initial relation  $xRy$ . The second step uses this pattern in order to check that the subject in  $xRy$  can be related to the object in  $yRz$ . In query language we denote this pattern as [ $x R z$ ]. In basic terms, we used this pattern to detect if it is possible to find evidence in the Web of a concrete relationship between two specific individuals.

**? (Object, Subject):** Given two individuals  $x$  and  $y$ , with this pattern we try to find verb phrases that represent relationships between them.

*Inverse Object Property:* Given a relation  $xRy$ , we can search for an inverse relationship by looking for a verb that relates  $y$  to  $x$ . In query language we denoted this pattern as [ $y * x$ ] OR [ $y * * x$ ] OR [ $y * * * x$ ]. Note that the wildcard (\*), which is used in the query in order to force the proximity (up to three words) between the subject and the object, has to correspond to a verb phrase.

### 3.2 Extraction

Applying the described patterns over a given relation and querying the resulting expression over a Web search engine, we are able to retrieve Web resources including one or several matchings  $\langle 10 \rangle$ . The text is analysed in order to discover and extract the semantic evidences that would lead to the definition or not of the corresponding

axiom. In order to minimize noise and ambiguity of textual processing, we imposed some syntactic boundaries. In general, we decided to get rid of interrogative sentences, cardinal candidates, candidates that are superclasses of the initial object, and verbs in future or gerund tense. Words between parentheses are deleted. If a text piece fulfills the conditions and match with the particular pattern, then the evidence is extracted <13> for further analysis.

### 3.3 Evaluating Extractions and Axiom Definition

Applying the mentioned constraints to avoid natural language problems does not mean that the observations extracted for a certain relation are strong enough to grant the definition of an axiom. The next step consists on deciding which of the extractions are reliable and sufficiently related to the property. To perform this selection process we performed a Web-based statistical analysis relying on co-occurrence measures computed directly from Web search engines <19>. Co-occurrence measures are based on distributional hypothesis claiming that words that occur in the same context tend to have similar meanings.

Several scores have been proposed in the past to compute Web-scale statistics, adapting the notion of collocation and mutual information (computed as the probability of joint appearance of concepts in a corpus). Applied to the Web, probabilities ( $p$ ) are estimated as the hit count (*hits*) returned by a Web search engine when querying the term. We used the Symmetric Conditional Probability (SCP) [13] computed from the Web (1) to evaluate the relatedness between two entities.

$$SCP(x,y) = \frac{p(x,y)^2}{p(x)*p(y)} \approx \frac{\left(\frac{\text{hits}("x \text{ AND } y")}{\text{total\_web\_sites}}\right)^2}{\frac{\text{hits}("x")}{\text{total\_web\_sites}} * \frac{\text{hits}("y")}{\text{total\_web\_sites}}} = \frac{\text{hits}("x \text{ AND } y")^2}{\text{hits}("x") * \text{hits}("y")} \quad (1)$$

To evaluate the relevance of the extracted terms, we need to measure the co-occurrence of the relation and the extracted evidence. In order to construct queries that approximate term appearance probabilities, we consider a *discriminator phrase*, which represents the  $xRy$  relation. We call *instance* to the expression created with the extracted term plus the relation that links it to another individual, and *discriminator phrase without instance* to the expression that represents the relation without the extracted *instance*. So, when evaluating if the relation  $xRy$  “Spain borders France” is functional when a new object “Portugal” is retrieved, we can denote from (1):

**p(x,y) - Discriminator Phrase:** “Spain” AND “borders” AND “Portugal”

**p(x) - Instance:** “borders” AND “Portugal”

**p(y) - Discriminator Phrase without Instance:** “Spain” AND “borders”

The Web-based statistical evaluation process for the reflexive, symmetric and transitive (in its second step) properties implies the assessment of the relative relatedness (computed by means of the SCP) of the extracted evidences with respect to the initial relation. If the SCP of the extracted evidence is higher than a threshold, we may conclude that the relation is reflexive, symmetric or transitive (in its second step) <21, 22>. For the functional, inverse, inverse functional and transitive (in its first step)

object properties, if the set of extractions contains an item with a co-occurrence higher than the threshold then we may conclude that the extractions are robust for the relation. Depending on these observations, an axiom may be fulfilled or not. In the functional and inverse functional case, if the extractions are robust then we may conclude that the relation is not functional or inverse functional <26, 27>. For the inverse axiom, we may conclude that an inverse relation has been found. In the first step of the transitivity test, we conclude that a good candidate to check for the transitivity in the second step has been found <21,22>.

## 4 Empirical Evaluation

Due to the difficulty to carry out automatic evaluations of automatically acquired knowledge, we based them in a manual approach in which a human determines if a relation fulfills a certain axiom. A number of positive and negative relation examples of each of the studied axioms are presented, along with the results given by our methodology. In the tables shown in this section the following abbreviations are used:

- SCP O/P: SCP values of the original relation and those obtained for the best extraction from the constructed pattern.
- D: Decision of the algorithm concerning the evaluated axiom relation.
- HD: Decision of a human concerning the evaluated axiom relation.
- NC: Number of extractions for a given axiom and relation.
- BES: Best extraction selected (the one with the highest SCP value), when applicable.

The examples used to evaluate the properties have been selected from different domains, in order to show the genericity of the proposed methodology. In all cases the threshold used to select or discard an extraction has been empirically set to  $1E-3$ . We used two Web search engines and the maximum number of Web pages retrieved for each pattern has been set to 500 per Web search engine.

**Symmetric Object Property.** The evaluation of the Symmetry of an Object Property has been tested with 10 relations, presented in Table 1 (5 of them are symmetric and 5 of them are not). The algorithm agrees in 9 out of 10 cases with the human evaluation.

**Table 1.** Test of Symmetry for several Object Properties

Original relation	Pattern	SCP O	SCP P	D	HD
Spain borders France	France borders Spain	0.22	0.22	Y	Y
Obesity is associated with cancer	Cancer is associated with obesity	0.09	0.09	Y	Y
Microsoft competes with Google	Google competes with Microsoft	0.29	0.29	Y	Y
Angelina Jolie is married to Brad Pitt	Brad Pitt is married to Angelina Jolie	0.34	0.34	Y	Y
Cigarettes are linked to cancers	Cancers are linked to cigarettes	0.001	0.001	Y	Y
Deep Blue defeated Gary Kasparov	Gary Kasparov defeated Deep Blue	0.004	0.004	<b>Y</b>	<b>N</b>
Heavy smokers get lung cancer	Lung cancer get heavy smokers	6.5E-4	-	N	N
Michael Jackson was born in Gary	Gary was born in Michael Jackson	0.006	-	N	N
Sociology is the study of society	Society is the study of Sociology	0.009	-	N	N
Madrid is located in Spain	Spain is located in Madrid	0.03	-	N	N

Note that, in this case, SCP O and SCP P coincide as the two individuals and the relationship are the same in the original relation  $xRy$  and in the relation  $yRx$ . In our test, 4 negative examples didn't obtain any result with the given pattern.

From the "Deep Blue defeated Gary Kasparov" example notice that, as a matter of fact, Gary Kasparov defeated Deep Blue but in a different match. This illustrates a limitation of the automatic process.

**Reflexive Object Property.** The evaluation of the reflexivity of an Object Property has been tested with 10 relations, presented in Table 2 (5 of them are reflexive and 5 of them aren't). A relation is considered reflexive if the SCP score of the pattern exceeds  $1E-3$ . The algorithm agrees in 8/10 cases with the human evaluation.

In the evaluated examples, the algorithm found evidences for "people have itself" and "stress affects itself". The correlations of "itself" with their respective subjects are high, but that doesn't imply they are correct. For these two cases, the human expert evaluated the reflexivity as wrong.

**Table 2.** Test of Reflexivity for several Object Properties

Original relation	Pattern relation	SCP O	SCP P	D	HD
Nature regulates CO2	Nature regulates itself	0.04	0.23	Y	Y
Market regulates capital production	Market regulates itself	2.88E-3	0.14	Y	Y
Fire consumes wood	Fire consumes itself	0.15	0.09	Y	Y
Science studies Nature	Science studies itself	0.07	0.03	Y	Y
Immune system attacks viruses	Immune system attacks itself	0.015	0.0048	Y	Y
Testicular cancer has good prognosis	Testicular cancer has itself	1.1E-4	3.6E-5	N	N
People have colon cancer	People have itself	0.002	0.16	Y	N
Pressure sensor measures water pressure	Pressure sensor measures itself	8.4E-3	1.9E-4	N	N
Hysterectomy reduces cancer	Hysterectomy reduces itself	0.012	7.2E-4	N	N
Stress affects cancer	Stress affects itself	0.05	0.046	Y	N

**Functional Object Property.** The evaluation of the Functionality of an Object Property has been tested with 10 relations, presented in Table 3 (5 of them are functional and 5 of them are not). The algorithm agrees in 8/10 cases with the human evaluation. The table shows the best extraction found with the relation(subject,?) pattern. The algorithm takes the relation to be functional if the best extraction has an SCP lower than the established  $1E-3$  threshold.

**Table 3.** Test of Functionality for several Object Properties

Original relation	NC	BES	SCP O	SCP P	D	HD
Brasilia is the capital of Brazil	4	country	0.004	3.7E-4	Y	Y
Cadmium exposure is associated with prostate cancer	1	reduced pulmonary function	1.28E-4	1.3E-4	Y	N
Da Vinci Code was written by Dan Brown	1	Bart D. Ehrman	0.009	6.5E-4	Y	Y
Comet Halley is named for Edmund Halley	0	-	0.019	-	Y	Y
Microsoft competes with Google	42	Apple	0.29	0.23	N	N
Brazil borders Bolivia	8	Peru	0.14	0.20	N	N
Michael Jackson was born in Gary	3	August	0.005	0.006	N	Y
Raloxifene reduces breast cancer	32	osteoporotic fractures	9.4E-4	0.006	N	N
Spain borders France	3	Portugal	0.229	0.25	N	N
Chichen Itza is known as El Castillo	2	La Iglesia	0.008	8.25E-5	Y	Y

It can be observed that the functionality test presents some limitations. The original relation “Cadmium exposure is associated with prostate cancer” scored a SCP of  $1.28E-4$  and the BES “reduced pulmonary function” scored higher than the original relation with  $1.3E-4$ , but the algorithm considered the relation as not functional. For this case the expert decided that the relation with the best extraction is in fact valid.

Regarding the relation “Michael Jackson was born in Gary”, the relation “was born in” may be related to either time or place giving unexpected results. The expert considered the relation as functional (giving an approach of “place” only) while the algorithm considered the contrary.

**Inverse Object Property.** The discovery of Inverse relations for an Object Property has been tested with 10 relations, presented in Table 4. The table shows the best extraction (i.e., the best candidate to inverse relation) found with the  $?(object, subject)$  pattern. The algorithm takes the relation to have an inverse if the best extraction has an SCP (SCP P column) higher than the established  $1E-3$  threshold.

**Table 4.** Discovery of Inverse relations for several Object Properties

Original relation	NC	BES	SCP O	SCP P	D	H D
Christopher Columbus discovered America	5	was discovered by	0.004	0.017	Y	Y
Market regulates capitalist production	1	is attached to	$1.8E-3$	$5.8E-5$	N	N
Global warming is caused by pollution	6	causes	0.11	0.06	Y	Y
Cigarette smoking is associated with lung cancer	1	is caused by	0.007	0.002	Y	Y
Sensors contain mechanical parts	1	required for	$4.7E-3$	$7.1E-4$	N	N
Viruses are associated with liver cancer	2	caused by	$5.2E-3$	0.003	Y	Y
Kasparov defeated Deep Blue	16	beaten by	0.003	0.001	Y	Y
Stress affects asthma	3	is linked to	0.02	0.025	Y	Y
Meningococcal disease is caused by Neisseria Meningitides	1	causes	0.03	0.002	Y	Y
Charles Darwin proposed evolution theory	8	proposed by	0.002	0.001	Y	Y

The examples show that it is possible to extract incomplete verb phrases. For example, we found evidences for “liver cancer caused by viruses”, whereas “is caused by” would be a more complete verb phrase for that relation.

**Inverse Functional Object Property.** The Inverse Functionality of an Object Property has been tested with 5 of the “inverse” relations shown in the previous table. Table 5 shows the best extraction found with the relation(subject,?) pattern. The algorithm

**Table 5.** Test of the Inverse Functionality of some Object Properties

Original relation	NC	BES	SCP O	SCP P	D	HD
Pollution causes global warming	76	acid rain	0.066	0.013	N	N
Asthma is linked to stress	67	allergies	0.026	0.108	N	N
Neisseria Meningitides causes meningococcal disease	2	inflammation	0.0022	$3.08E-5$	Y	Y
Evolution theory proposed by Charles Darwin	1	Darwin	0.001	0.001	N	Y
Lung cancer is caused by cigarette smoking	57	second hand smoke	0.003	0.0028	N	N



takes the original relation to be inverse functional if the best extraction has an SCP (SCP P column) higher than the established  $1E-3$  threshold. The algorithm agrees with the human result in 4/5 cases. In the example (shown in the following table) that states “Evolution theory proposed by Charles Darwin”, notice that a selected extraction is “Darwin”. Both expressions refer to the same individual, but with the algorithm proposed this fact is not taken into account concluding in a wrong decision.

**Transitive Object Property.** To test the Transitivity of an Object Property we have followed two steps. Given a relation  $xRy$ , in the first step we look for individuals related to  $y$  via  $R$ , using the pattern  $R(y, ?)$ . In the second step, we take the best extraction in each of the cases ( $z$ ), and we try to evaluate whether  $xRz$  holds.

Table 6a shows the five relationships used in this test, along with the best extraction found in each case. An extraction is taken as correct by the algorithm if its SCP is above the  $1E-3$  threshold. From the first four relations given in Table 6a, we evaluate in Table 6b the best extraction ( $z$ ) to see if  $R(x, z)$  holds. If it does, the algorithm takes the relation to be transitive.

**Tables 6 a and b.** Two steps in the test of Transitivity for some Object Properties

Original relation	NC	BES	SCP O	SCP P	D	HD
Global warming is caused by pollution	54	Nature	0.11	0.05	Y	Y
Berlin is located in Germany	12	Europe	0.04	0.07	Y	Y
Pollution affects global warming	46	Weather	0.038	0.05	Y	Y
Microsoft competes with Apple	30	Google	0.29	0.18	Y	Y
Arsenic is associated with kidney cancer	5	Obesity	8.1E-5	4.3E-5	N	Y

Relation	SCP O	SCP P	D	HD
Global warming is caused by Nature	0.11	0.05	Y	Y
Berlin is located in Europe	0.04	0.02	Y	Y
Pollution affects weather	0.038	0.02	Y	Y
Microsoft competes with Google	0.29	0.31	Y	Y

Considering that the evaluation of the transitive property in our approach includes two steps, the possibility to find transitivity for a relation is difficult. However, if a transitivity property is found using this approach that is due to the fact that the evaluated relatedness is certainly strong, as in the relations “Berlin is located in Europe” or “Microsoft competes with Google”.

## 5 Conclusion and Future Work

Axiomatizing ontologies by hand is a hard and time-consuming task which involves ontology engineers and domain experts. The design of automatic axiomatization solutions is fundamental to contribute to the success of the Semantic Web and many other fields in Computer Science.

Our approach is a step forward to fill this gap, by learning object properties axioms for non-taxonomic relations using the Web as corpus, shallow linguistic techniques and Web-scale statistics. The Web can be exploited to aid in the Ontology Learning

process by using suitable queries in order to retrieve valuable information. The patterns used in this work are totally generic and can be easily applied to any Web search engine. Besides, the proposed method is domain-independent.

However, being completely automatic and unsupervised, the proposed algorithm presents several limitations, that can lead to the definition of lines of future work. Some of these shortcomings are the following:

- At the moment we study only the axioms of a particular relationship between two individuals,  $xRy$ , e.g. to test the symmetry, we check whether  $yRx$  holds. We don't cover the study of the mathematical properties of the  $R$  relation in general; for example, to guarantee the symmetry of  $R$ , we should check whether, for all  $x,y$  such that  $xRy$  holds,  $yRx$  also holds.
- Regarding the reflexive property, the search for " $xR$  itself" may be too restrictive and hard to find explicitly. We can add to the query the words "may" and "can" in order to extend the reflexive property discovery in the form of *Subject may/can Relation itself*.
- The transitive property is checked only with a particular instance. Note that "Italy borders Switzerland" will be qualified as transitive depending on the country which is the best extraction for the "Switzerland borders" pattern (giving different results e.g. for France and Germany, as one borders Italy and the other doesn't). Thus, we should at least check whether all the neighbours of Switzerland are also neighbours of Italy.

It is also necessary to study more deeply the effect of many natural language phenomena such as synonyms and polisemy in order to provide better results in certain domains in which these ambiguities are common. This may lead to further research in composing a more concrete set of patterns or adding more constraints in their extraction and selection. The empirical tests also show that in some cases the semantic relatedness acquired from the extractions has a similar value than the one of the original relation presented in the ontology. Assuming that the relations in the ontology are correct, we can use the SCP of the relations as a criterion select or discard observation instead of using a predefined threshold.

## Acknowledgements

Luis Miguel Del Vasto has been supported by a grant of Fundación Carolina. This work has been supported by the Universitat Rovira i Virgili (2009AIRE-04), the Spanish Ministry of Science and Innovation (DAMASK project, Data mining algorithms with semantic knowledge, TIN2009-11005) and the Spanish Government (PlanE, Spanish Economy and Employment Stimulation Plan).

## References

1. Sánchez, D.: Domain Ontology Learning from the Web. VDM Verlag Dr. Mueller (2008)
2. Cimiano, P.: Ontology Learning and Population from Text. Springer Science, 9–10 (2006)
3. Euzenat, J.M.: Research Challenges and perspectives of the Semantic Web. In: Report of the EU-NSF strategic workshop, Sophia-Antipolis, France, pp. 86–88 (2001)

4. Volker, J., Haase, P., Hitzler, P.: Learning Expressive Ontologies. In: Conference on Ontology Learning and Population, Germany, pp. 45–69 (2008)
5. Sánchez, D., Moreno, A.: Pattern-based automatic taxonomy learning from the Web. *AI Communications* 21(1), 27–48 (2008)
6. Shamsfard, M., Abdollahzadeh, A.: Learning Ontologies from Natural Language Texts. *International Journal of Human-Computer Studies* 60, 17–63 (2004)
7. Rudolph, S., Volker, J., Hitzler, P.: Supporting Lexical Ontology Learning by Relational Exploration. In: Priss, U., Polovina, S., Hill, R. (eds.) *ICCS 2007. LNCS (LNAI)*, vol. 4604, pp. 488–491. Springer, Heidelberg (2007)
8. Volker, J., Hitzler, P., Cimiano, P.: Acquisition of OWL DL Axioms from Lexical Resources. In: 4th European Conference on The Semantic Web, Germany, pp. 670–685 (2007)
9. Hearts, M.: Automatic Acquisition of Hyponyms from Large Text Corpora. In: 14th Conference on Computational linguistics, France, vol. 2, pp. 539–545 (1992)
10. Ritter, A., Downey, D., Soderland, S., Etzioni, O.: It’s a Contradiction-No it’s not: A Case study using Functional Relations. In: Conference on Empirical Methods in Natural Language Processing. University of Washington, USA, pp. 11–20 (2008)
11. Volker, J., Vrande, D., Sure, Y., Hotho, A.: Learning Disjointness. In: 4th European Conference on The Semantic Web, Germany, pp. 175–189 (2007)
12. Schutz, A., Buitelaar, P.: RelExt: A Tool for Relation Extraction from Text in Ontology Extension. In: Gil, Y., Motta, E., Benjamins, V.R., Musen, M.A. (eds.) *ISWC 2005. LNCS*, vol. 3729, pp. 593–606. Springer, Heidelberg (2005)
13. Dias, G., Santos, C., Cleuzious, G.: Automatic Knowledge Representation using a Graph-based Algorithm for Language-Independent Lexical Chaining. In: *Workshop on Information Extraction Beyond The Document*, Sydney, pp. 36–47 (2006)
14. Sánchez, D., Moreno, M.: Learning non-taxonomic relationships from web documents for domain ontology construction. *Data and Knowledge Engineering* 63(3), 600–623 (2008)
15. Rudolph, T.: Exploring relational structures via FLE. In: Wolff, K.E., Pfeiffer, H.D., Delugach, H.S. (eds.) *Conceptual Structures at Work*, pp. 196–212 (2004)

# An Improved Piecewise Aggregate Approximation Based on Statistical Features for Time Series Mining

Chonghui Guo, Hailin Li, and Donghua Pan

Institute of Systems Engineering,  
Dalian University of Technology, Dalian, 116024, China  
guochonghui@tsinghua.org.cn, hailin@mail.dlut.edu.cn

**Abstract.** Piecewise Aggregate Approximation (PAA) is a very simple dimensionality reduction method for time series mining. It minimizes dimensionality by the mean values of equal sized frames, which misses some important information and sometimes causes inaccurate results in time series mining. In this paper, we propose an improved PAA, which is based on statistical features including a mean-based feature and variance-based feature. We propose two versions of the improved PAA which have the same preciseness except for the different CPU time cost. Meanwhile, we also provide theoretical analysis for their feasibility and prove that our method guarantees no false dismissals. Experimental results demonstrate that the improved PAA has better tightness of lower bound and more powerful pruning ability.

**Keywords:** Piecewise aggregate approximation, statistical feature, time series, dimensionality reduction.

## 1 Introduction

Time series is a kind of the data with time property. In most cases, efficient and accurate similarity search in time series dataset, including indexing, pattern discovery and association rule, is a very important task for knowledge and information mining. Time series is a kind of data with high dimensionality. It is difficult and inefficient to directly mine time series without dimensionality reduction. Therefore, we should use some methods to reduce the dimensionality before mining.

There has been much work in dimensionality reduction for time series. Some popular methods include discrete fourier transform (DFT) [1], discrete wavelet transform (DWT) [2], singular value decomposition (SVD) [3], symbolic aggregate approximation (SAX) [4, 5] based on piecewise aggregate approximation (PAA) [6], adaptive piecewise constant approximation (APCA) [7]. However, PAA is one of the simplest dimensionality reductions. It reduces dimensionality by the mean of equal sized segments, which also causes the missing of some important features for some kinds of time series datasets. Some people have

found the problem and considered the mean values with slope values to improve PAA [10]. What they have done add some valuable information to time series mining and make PAA more perfect.

In this paper we use two important statistic features including the means values and variance values as key factors to improve the original PAA. In fact, besides considering the mean values of the equal sized segments, variance values are also important for time series mining. The improved PAA methods could explain the distribution of the points of each equal sized segment, which provides more information for time series data mining. In this work, besides the theoretical analysis, we also make some experiments to show the performance of the improved PAA methods.

The rest of the paper is organized as follows. Section 2 introduces the original PAA and discusses the existing problem. In section 3 we propose two versions of the new approach. Section 4 shows the experiments and empirical comparisons of our method with completing techniques. In section 5 we offer conclusions and suggestions for the future work.

## 2 Piecewise Aggregate Approximation

Similarity search is a very important task for time series data mining. Since time series is a kind of data with high dimensionality, we always reduce the dimensionality in advance. It means that we should transform data from high space into low space. After dimensionality reduction we can derive a new distance function which is applied to the low space. As proved in reference [1], the function should be a lower bounding measure, which guarantees no false dismissals. In other words, the new function in the low space should underestimate the true distance measure. Conventionally, there are two available true distance measures, Euclidean distance and dynamic time warping (DTW) [8]. In this work, we regard Euclidean distance as the true distance measure.

Suppose we have two time series  $Q$  and  $C$  of dimension  $m$  in original space. Euclidean measure  $D(Q, C)$  is the true distance measure function. Another two time series  $Q'$  and  $C'$  are the new version and  $LB\_D(Q', C')$  is the new distance function in the low space. The two functions must satisfy

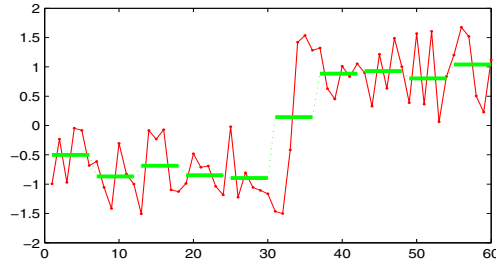
$$D(Q, C) \geq LB\_D(Q', C'). \quad (1)$$

PAA is a lower bounding measure function, which is proposed by Lin et al [6]. It is also extended for the better application. Symbolic aggregate approximation (SAX) is used to condense the time series and transform them into a symbolic string, which has wide use in many field. Moreover, since the mature technique PAA is very simple, the extended SAX [9] is also very popular.

A time series  $Q$  of length  $m$  can be transformed into another form by a vector  $\bar{Q} = \bar{q}_1, \bar{q}_2, \dots, \bar{q}_w$  in a  $w$ -dimensional space. Then the  $i$ th element of  $\bar{Q}$  can be calculated by

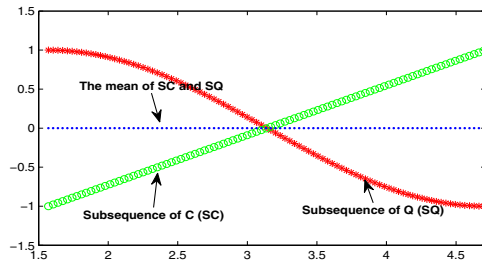
$$\bar{q}_i = \frac{1}{k} \sum_{j=k(i-1)+1}^{k*i} q_j, \quad \text{where } k = \frac{m}{w}. \quad (2)$$

Actually, when original time series in  $m$  dimensional space is transformed into another one in  $w$  dimensional space, the time series is divided into  $w$  segments with equal size. The mean value of the data within a segment is obtained and a vector of these values is the final representation in the new space, as shown in Fig.1.



**Fig. 1.** PAA representation is obtained by transformation of space. In this case, a time series of length 60 is reduced to 10 dimensions.

Since PAA is a method based on the mean of each segment, it misses some valuable features for some kinds of time series. From Fig.1 we know that PAA only approximates the whole trend and loses the variance features in each segment. Moreover, PAA causes the inaccurate result. As shown in Fig.2, the two different subsequence of time series,  $SQ$  and  $SC$ , have the same mean. Therefore, the original PAA may produce the inaccurate result for some time series.



**Fig. 2.** Two different subsequences of time series have the same mean value

For the above case, although the two different subsequences have the same means, they have the different variance. We can use another statistical feature

(variance) to approximate time series because variance can provide some additional information. Therefore, our new approach is based on the mean and the variance of time series for similarity search.

### 3 An Improved Piecewise Aggregate Approximation

Although PAA is one of the best techniques to reduce the dimensionality for time series, it also has some disadvantages. We propose another two versions of new approach which is combination of two statistical features, mean and variance. One version of the new approach is the linear combination of two distance measure functions, which are mean distance measure function and variance distance measure function. We call it Linear Statistical Feature based Piecewise Aggregate Approximation (LSF\_PAA). The other is the square root of the sum of the two distance measure functions. We call it Square root Statistical Feature based Piecewise Aggregate Approximation (SSF\_PAA).

#### 3.1 Linear Statistical Feature Based Piecewise Aggregate Approximation

In section 2, we discuss how to get the mean of each segment. All the mean values of time series are the elements of a new vector in a low space. Likewise, in the new space we can get the variance values of time series.

A time series  $Q$  of length  $m$  is represented in a  $w$  dimensional space by a variance vector  $\hat{Q} = \hat{q}_1, \hat{q}_2, \dots, \hat{q}_w$ . The  $i$ th element of  $Q$  is calculated by

$$\hat{q}_i = \sqrt{\frac{1}{k} \sum_{j=k(i-1)+1}^{k*i} (q_j - \bar{q}_i)^2}, \quad \text{where } k = \frac{m}{w}, \tag{3}$$

where  $\bar{q}_i$  is the mean value of the  $i$ th segment in  $w$  dimensional space.

The two statistical features, the mean value  $\bar{q}_i$  and the variance value  $\hat{q}_i$ , describe the time series in low dimensional space more clearly. Especially the variance can provide additional information and allow the distance measure function in low dimensional space reflect the distribution of the points of every segment. To overcome the disadvantages of PAA, we propose the Linear Statistical Feature based Piecewise Aggregate Approximation (LSF\_PAA), which is linear combination of the two distance measure functions.

Given two time series  $Q$  and  $C$  of length  $m$ . We can view them as 2 vectors in original space. They can be transformed into 4 new vectors in a  $w$  dimensional space,  $\bar{Q}$ ,  $\bar{C}$ ,  $\hat{Q}$  and  $\hat{C}$ . Now we have some distance measure functions about the above 4 new vectors plus 2 original vectors.

The Euclidean distance measure is regarded as the true distance of time series, which is defined by

$$D(Q, C) = \sqrt{\sum_{i=1}^m (q_i - c_i)^2}. \tag{4}$$

The mean distance measure function of PAA is

$$\bar{D}(\bar{Q}, \bar{C}) = \sqrt{k \sum_{i=1}^w (\bar{q}_i - \bar{c}_i)^2}. \quad (5)$$

Similarly, we have a variance distance

$$\hat{D}(\hat{Q}, \hat{C}) = \sqrt{k \sum_{i=1}^w (\hat{q}_i - \hat{c}_i)^2}. \quad (6)$$

Finally, we obtain the LSF\_PAA measure function,

$$LB\_DL(Q, C) = \bar{D}(\bar{Q}, \bar{C}) + \mu * \hat{D}(\hat{Q}, \hat{C}), \quad (7)$$

where  $\mu \in [0, 1]$ . It is a lower bounding function measure.

**Proposition 1:** If there is  $\mu \geq \frac{\sqrt{\bar{D}^2 + \hat{D}^2} - \bar{D}}{\hat{D}}$ , then we have  $D(Q, C) \geq LB\_DL(Q, C)$ , i.e.  $D(Q, C) \geq \bar{D}(\bar{Q}, \bar{C}) + \mu * \hat{D}(\hat{Q}, \hat{C})$ .

**Proof:** We denote  $q_i = \bar{q}_i + \Delta q_i$  and  $c_i = \bar{c}_i + \Delta c_i$ . For the simple proof, we let  $w = 1$ , then  $q_i = \bar{q} + \Delta q_i$ ,  $c_i = \bar{c} + \Delta c_i$ , where  $\bar{q}$  and  $\bar{c}$  are the mean of Q and C respectively.

$$\begin{aligned} \sum_{i=1}^m (q_i - c_i)^2 &= \sum_{i=1}^m ((\bar{q} + \Delta q_i) - (\bar{c} + \Delta c_i))^2 \\ &= \sum_{i=1}^m ((\bar{q} - \bar{c}) + (\Delta q_i - \Delta c_i))^2 \\ &= m(\bar{q} - \bar{c})^2 + 2(\bar{q} - \bar{c}) \sum_{i=1}^m (\Delta q_i - \Delta c_i) + \sum_{i=1}^m (\Delta q_i - \Delta c_i)^2. \end{aligned}$$

Moreover, we have

$$\begin{aligned} (\bar{q} - \bar{c}) \sum_{i=1}^m (\Delta q_i - \Delta c_i) &= (\bar{q} - \bar{c}) \sum_{i=1}^m ((q_i - c_i) - (\bar{q} - \bar{c})) \\ &= (\bar{q} - \bar{c}) \sum_{i=1}^m (q_i - c_i) - m(\bar{q} - \bar{c})^2 \\ &= (\bar{q} - \bar{c}) m \left( \frac{1}{m} \sum_{i=1}^m (q_i - c_i) \right) - m(\bar{q} - \bar{c})^2 \\ &= (\bar{q} - \bar{c}) m(\bar{q} - \bar{c}) - m(\bar{q} - \bar{c})^2 \\ &= 0. \end{aligned}$$

Therefore,

$$D^2(Q, C) = m(\bar{q} - \bar{c})^2 + \sum_{i=1}^m (\Delta q_i - \Delta c_i)^2. \quad (8)$$



Notice that,

$$\sum_{i=1}^m (\Delta q_i - \Delta c_i)^2 = \sum_{i=1}^m \Delta q_i^2 + \Delta c_i^2 - 2\Delta q_i \Delta c_i \tag{9}$$

and

$$\begin{aligned} \hat{D}^2(\hat{Q} - \hat{C}) &= m(\hat{q} - \hat{c})^2 \tag{10} \\ &= m\left(\sqrt{\frac{1}{m} \sum_{i=1}^m \Delta q_i^2} - \sqrt{\frac{1}{m} \sum_{j=1}^m \Delta c_j^2}\right)^2 \\ &= \sum_{i=1}^m \Delta q_i^2 + \sum_{j=1}^m \Delta c_j^2 - 2\sqrt{\sum_{i=1}^m \Delta q_i^2 \sum_{j=1}^m \Delta c_j^2}. \end{aligned}$$

Therefore, by Cauchy-Schwarz inequality, we have

$$\sqrt{\sum_{i=1}^m \Delta q_i^2 \sum_{j=1}^m \Delta c_j^2} \geq \sum_{i=1}^m \Delta q_i \Delta c_i \tag{11}$$

and

$$\sum_{i=1}^m (\Delta q_i - \Delta c_i)^2 \geq \hat{D}^2(\hat{Q} - \hat{C}). \tag{12}$$

Suppose we have

$$D^2(Q, C) \geq (\bar{D}(\bar{Q}, \bar{C}) + \mu \hat{D}(\hat{Q}, \hat{C}))^2, \quad \mu \in [0, 1]. \tag{13}$$

From formula (8) and (12), the inequality (13) becomes

$$\hat{D}^2(\hat{Q}, \hat{C}) \geq 2\mu \bar{D}(\bar{Q}, \bar{C}) \hat{D}(\hat{Q}, \hat{C}) + \mu^2 \hat{D}^2(\hat{Q}, \hat{C}).$$

Since there is  $\hat{D}(\hat{Q}, \hat{C}) \geq 0$ , we have  $\hat{D}(\hat{Q}, \hat{C}) \geq 2\mu \bar{D}(\bar{Q}, \bar{C}) + \mu^2 \hat{D}(\hat{Q}, \hat{C})$ . Through solving this one-variable quadratic inequality with respect to  $\mu$ , the maximum value of the variable  $\mu = \frac{\sqrt{\bar{D}^2 + \hat{D}^2} - \bar{D}}{\hat{D}} \in [0, 1]$  is obtained.

The proof is finished.

By far we have taken a theoretical analysis which proves that LSF\_PAA is a lower bounding measure. Moreover, it is a better tightness of lower bound, i.e.

$$D(Q, C) \geq LB\_DL(Q, C) \geq \bar{D}(\bar{Q}, \bar{C}). \tag{14}$$

In fact, LSF\_PAA is comprised of the measure function of PAA and the variance measure function. The contribution is  $\mu * \hat{D}(\hat{Q}, \hat{C})$ . Therefore, from the mathematical analysis we also know LSF\_PAA is tighter than original PAA.

### 3.2 Square Root Statistical Feature Based Piecewise Aggregate Approximation

Another version of our new approach is the square root of the sum of the mean and the variance distance measure functions. We call it Square root Statistical Feature based Piecewise Aggregate Approximation (SSF\_PAA). Just as illustrated in subsection 3.1, we have the mean distance measure and the variance distance measure. The SSF\_PAA distance measure function is

$$LB\_DS(Q, C) = \sqrt{\bar{D}^2(\bar{Q}, \bar{C}) + \hat{D}^2(\hat{Q}, \hat{C})}. \quad (15)$$

**Proposition 2:** If there is  $LB\_DS(Q, C) = \sqrt{\bar{D}^2(\bar{Q}, \bar{C}) + \hat{D}^2(\hat{Q}, \hat{C})}$ , then we have  $D(Q, C) \geq LB\_DS(Q, C)$ , i.e.  $LB\_DS(Q, C) \geq \sqrt{\bar{D}^2(\bar{Q}, \bar{C}) + \hat{D}^2(\hat{Q}, \hat{C})}$

**Proof:** From formula (8), (9), (10) and (11), we derive  $\sum_{i=1}^m (q_i - c_i)^2 \geq m(\bar{q} - \bar{c})^2 + \hat{D}^2(\hat{Q}, \hat{C})$ , i.e.

$$D^2(Q, C) \geq \bar{D}^2(\bar{Q}, \bar{C}) + \hat{D}^2(\hat{Q}, \hat{C}).$$

The proof of the SSF\_PAA is finished.

Through the above mathematical analysis, SSF\_PAA is also a lower bounding measure function, which can guarantee no false dismissals. Moreover, the lower bounding of this function is tighter than PAA too. It is also easy to obtain

$$D(Q, C) \geq LB\_DS(Q, C) \geq \bar{D}(\bar{Q}, \bar{C}). \quad (16)$$

### 3.3 Complexity Analysis of LSF\_PAA and SSF\_PAA

The above two versions of the new improved PAA have the same effectiveness to reduce dimensionality. Moreover, their tightness of lower bound is identical. However, their time consumption of similarity search is different.

A time series  $Q$  of length  $m$  is used to query its similar objects by brute-force searching method in time series dataset with  $n$  time series. In LSF\_PAA, it calculates the parameter value  $\mu$  in advance for each pair of time series, which totally costs  $O(mn)$ . However, since SSF\_PAA has no any parameter, it doesn't cost the time. Therefore, SSF\_PAA is faster than LSF\_PAA.

From the above formulas, we know that each version of the new approach has one more function than PAA, which causes more time consumption than PAA. The additional time of SSF\_PAA is used to calculate the variance value of each segment. This additional time is equal to the time consumption of the mean values of each segment. Thereby, the time complexity of SSF\_PAA is twice as that of PAA. In LSF\_PAA, because it needs to cost  $O(mn)$  to calculate the parameter and also needs to calculate the mean and variance value, its time consumption is much higher than PAA. In conclusion, the time consumption of SSF\_PAA is a little higher than PAA but lower than LSF\_PAA.

If we only consider the precision of the similarity search, we can choose the two versions. If we consider the productivity of search algorithm, the SSF\_PAA is a better choice.

## 4 Numerical Experiments

In this section we test our approach with a comprehensive set of experiments. The experimental data is from the UCI Dataset (Synthetic Control Chart Time Series) [11]. The purpose of the experiments is to concentrate on reducing the dimensionality without false dismissals. The comparison among PAA, SSF\_PAA and LSF\_PAA is based on the tightness of lower bounds, pruning power and implement system as introduced in reference [12].

### 4.1 Comparison of Lower Bounding Measures

From the above discussions, we can conclude that the tightness of the two versions of improved PAA is better than the original PAA. Now we empirically test whether the conclusion is true. Since the three versions of PAA are the approximation of the true distance, we directly call them estimated distance. We let  $T$  represent tightness and define it as the ratio of the estimated distance between two time series over the true distance between the same time series, i.e.  $T = D'(Q, C)/D(Q, C)$ , where  $D(Q, C)$  is a true distance measure and  $D'(Q, C)$  is the estimated distance measure including PAA, SSF\_PAA and LSF\_PAA.  $T$  is in the range  $[0, 1]$ , with the larger the better.

To experiment on the tightness of lower bound of the three versions, 600 time series of length 60 were tested. Each time series was condensed into a new vector of elements. We compare each time series to the other 599 and report  $T$  as average ratio from 179700 ( $600 \cdot 599 / 2$ ) comparisons we made. The tightness results of the three versions of PAA with regard to reduction ratio  $k = m/w$  are shown in Fig.3.

We found that the tightness of SSF\_PAA is the same to that of LSF\_PAA, which means the two version of improved PAA can produce same distance measure. Although they are identical, they are larger than the original PAA for each

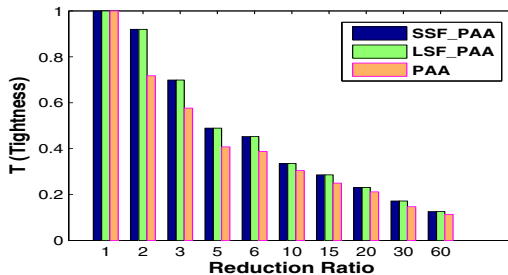


Fig. 3. The empirically estimated tightness of the three versions of PAA

reduction ratio. When the reduction ratio is equal to 1, it means the two versions correspond to the original PAA.

## 4.2 Comparison of Pruning Power

Pruning power indicates that the estimated measure function can decrease the number of time series which require full computation of true distance when indexing time series or similarity search. In other words, it is the fraction of the dataset that must not be examined before we can guarantee that we have found the nearest query. We define it as the ratio of the number of objects that do not require full computation of true distance over the number of object existing in the time series database, i.e.

$$P = \frac{\text{Number of objects that are not examined by true distance function}}{\text{Number of objects in database}} \quad (17)$$

We randomly extract 100 time series from the database and regard them as query time series. We let everyone search the similar time series with regard to the reduction ratio  $k = m/w$ . In Fig.4, the result of pruning power shows that the two version of the improved PAA also have the same ability to reject the time series which do not require full computation by the true distance measure. Fortunately, they are larger than the original PAA. Moreover, when the reduction ratio is 1, namely,  $w = 60$ , it means the two version of the improved PAA degenerate to the original PAA.

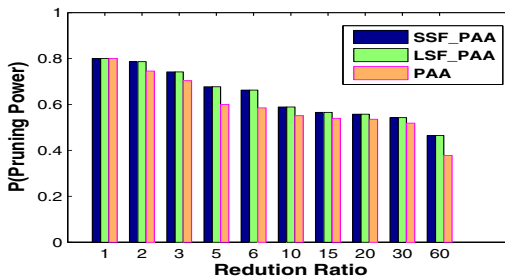
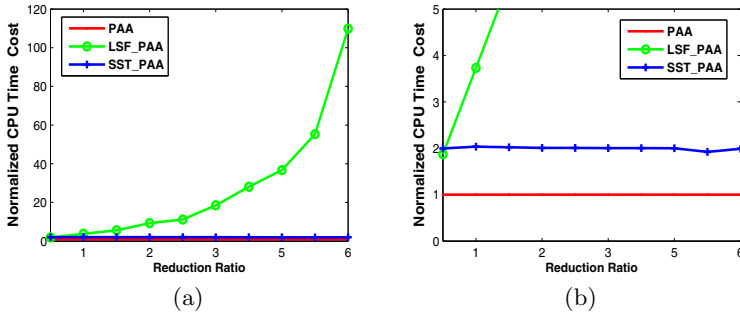


Fig. 4. The pruning power of the three version of PAA

## 4.3 Experiments on Implemented System

To evaluate the performance of the three versions of PAA, we test the normalized CPU cost. We define the normalized CPU cost as the ratio of average CPU time of the three version of PAA to query the time series over the average CPU time of PAA. It is easy to know that the normalized CPU time of PAA is equal to 1. We performed the experiments on Intel Core2 Duo 2.00GHZ processor with 2.00GB physical memory. We also experiment on the time series database and



**Fig. 5.** The Normalized CPU time cost of the three versions of PAA. (a) The original view of the normalized CPU time cost; (b) The zoom view of the normalized CPU time cost.

arbitrarily choose 150 time series as the query ones. We let them query in the database according to different reduction ratio  $k = m/w$ . The Fig.5(a) shows the result of the experiment.

There are at least three kinds of information hiding in Fig.5(b). The first is that the time consumption of LSF\_PAA is much too large. The second is that the time consumption of SSF\_PAA is approximately twice as that of PAA. We also can find that the normalized CPU time of LSF\_PAA is equal to that of SSF\_PAA when the reduction ratio  $k$  is equal to 1. In this case, it means the LSF\_PAA corresponds to SSF\_PAA, which is the last information. We also point out that the larger the size of the database is, the more time is cost by SSF\_PAA, letting alone the LSF\_PAA.

## 5 Conclusions

The main contribution of this paper is to propose two improved versions of the improved PAA in light of the precision. Meanwhile, we also provide their mathematical proof with regard to the lower bounding measure. Through considering the variance of each segment in the low space, we can let the new approach provide additional information and make new estimated distance measure have better tightness of lower bound. From the view of the accuracy of similarity search, the two versions of PAA are better than the original one. Although they cost more time, fortunately, SSF\_PAA's time consumption is only twice as that of the original PAA and its effectiveness is better. LSF\_PAA is not suitable for the large time series database because of its high time complexity. However, its idea is worth further researching in the future.

Considering the limitation of the improved approach, one of future work is to find a way to decrease time consumption. Since PAA has many extensions and is applied to various fields, we may extent the improved approach like the means of extended PAA.

## Acknowledgment

This work has been partly supported by the Natural Science Foundation of China under Grant No. 70871015 and the National High Technology Research and Development Program of China under Grant No. 2008AA04Z107.

## References

1. Faloutsos, C., Ranganathan, M., Manolopoulos, Y.: Fast subsequence matching in time series databases. In: Proceedings of the ACM SIGMOD International Conference on Management of Data, pp. 419–429 (1994)
2. Popivanov, I., Miller, R.J.: Similarity search over time-series data using wavelets. In: Proceedings of the 18th International Conference on Data Engineering, pp. 212–221 (2002)
3. Theodoridis, S., Koutroumbas, K.: Feature generation I: data transformation and dimensionality reduction. In: Pattern Recognition, 4th edn., pp. 323–409 (2009)
4. Lin, J., Keogh, E.: A symbolic representation of time series with implications for streaming algorithms. In: Proceedings of the 8th ACM SIGMOD Workshop on Research Issues in Data Mining and Knowledge Discovery, pp. 2–11 (2003)
5. Keogh, E., Lin, J., Fu, A.: Hot sax: efficiently finding the most unusual time series subsequence. In: Proceedings of the 5th IEEE International Conference on Data Mining, pp. 226–233 (2005)
6. Lin, J., Keogh, E., Wei, L., Lonardi, S.: Experiencing SAX: a novel symbolic representation of time series. *Data Mining and Knowledge Discovery* 15, 107–144 (2007)
7. Keogh, E., Chakrabarti, K., Mehrotra, S., Pazzani, M.: Locally adaptive dimensionality reduction for indexing large time series databases. In: Proceedings of the ACM SIGMOD Conference on Management of Data, pp. 151–162 (2001)
8. Rabiner, L., Juang, B.H.: Fundamentals of speech recognition. Prentice Hall, Englewood Cliffs (1993)
9. Lkhagva, B., Suzuki, Y., Kawagoe, K.: New time series data representation ESAX for financial applications. In: Proceedings of the 22nd International Conference on Data Engineering Workshops, pp. 115–120 (2006)
10. Hung, N.Q.V., Anh, D.T.: An improvement of PAA for dimensionality reduction in large time series databases. In: Proceedings of the 10th Pacific Rim International Conference on Artificial Intelligence: Trends in Artificial Intelligence, pp. 698–707 (2008)
11. Pham, D.T., Chan, A.B.: Control chart pattern recognition using a new type of self organizing neural network. *Proceedings of the Institution of Mechanical Engineers, Part I: Journal of Systems and Control Engineering* 212, 115–127 (1998)
12. Keogh, E., Ratanamahatana, C.A.: Exact indexing of dynamic time warping. *Knowledge and Information Systems* 7, 358–386 (2005)

# Incorporating Duration Information in Activity Recognition

Priyanka Chaurasia<sup>1</sup>, Bryan Scotney<sup>1</sup>, Sally McClean<sup>1</sup>,  
Shuai Zhang<sup>1</sup>, and Chris Nugent<sup>2</sup>

<sup>1</sup> School of Computing and Information Engineering  
University of Ulster, Coleraine, Northern Ireland  
chaurasia-p@email.ulster.ac.uk,  
{bw.scotney, si.mcclean, s.zhang}@ulster.ac.uk

<sup>2</sup> School of Computing and Mathematics  
University of Ulster, Newtownabbey, Northern Ireland  
cd.nugent@ulster.ac.uk

**Abstract.** Activity recognition has become a key issue in smart home environments. The problem involves learning high level activities from low level sensor data. Activity recognition can depend on several variables; one such variable is duration of engagement with sensorised items or duration of intervals between sensor activations that can provide useful information about personal behaviour. In this paper a probabilistic learning algorithm is proposed that incorporates episode, time and duration information to determine inhabitant identity and the activity being undertaken from low level sensor data. Our results verify that incorporating duration information consistently improves the accuracy.

**Keywords:** Activity recognition, smart homes, duration.

## 1 Introduction

The number of elderly people in the population is increasing [1], and many older people may present with degenerative disease that can affect their cognitive abilities. The affects can vary in severity [2]. A compromise between the need for constant personalised care and the burden on caregivers is required [3]. It has been hypothesised that many elder people can live an independent life and their stay at home can be extended by the aid of at-home assistance and health monitoring [4].

The increase in cost of healthcare and motivation to help the elderly to live an independent life necessitate movement towards an advanced technological solution that can be embedded into a person's home. There have been several approaches to such assistive living. After learning the inhabitant's pattern of activity, a system can prompt the patient, alert the caregiver and inform the health status of the patient. Systems such as Autominder, [1] assist the elderly by creating daily plans, tracking the execution and making decisions. Personal Cognitive Assistance was introduced in

[5] to help with different kinds of deficit problems. The well-being of a person can be monitored by observing the changes in their activity patterns over time [6]. The assumption is that people carry out activities in a habitual way. Recognising activities of daily living (ADLs) is a major issue in smart homes, and much research is conducted in this field to monitor the key ADLs such as ‘eating’, ‘washing’ and ‘managing basic needs’ [1],[7].

A smart home is a distributed environment in which a set of sensors and computational devices are embedded in objects of interest to create a network that can sense the environment and give information about the action at a low-level which is then passed to the higher-level for further processing. There are different kinds of sensor such as audiovisual sensors, passive infrared sensors and multimodal wearables [8]. Activities of daily living can be identified by studying user interactions with objects and appliances fitted with sensors. Sensor data provide information about the objects being used [9]. The goal of activity recognition is to learn the actions and goals of the inhabitant from the information obtained from the sequence of sensor activations. There are several challenges related to activity recognition in a smart home. Firstly, ADLs can be performed in different ways. Secondly, recognising a particular ADL can be complicated by people multitasking. Thirdly, ADL recognition must maintain personal privacy and should be unobtrusive [9]. Fourthly, sensor networks are unreliable: they may malfunction and consequently inhabitant interaction with the environment may not be correctly reported [10].

Sensor data are time stamped, so the activity duration can be defined using the intervals between sensor activations [11]. Duration is therefore one of the additional features provided by the sensors along with information about the objects being used. Each person may take different amounts of time to complete a particular activity, and also activities typically have different durations. For example, making tea takes less time than making a meal.

Our objective here is to identify both the person and the activity. This can further contribute to prompting the person, giving feedback to the caregivers and reporting the health status of the person. This paper explores the use of duration for person identification and activity recognition. The structure of this paper is as follows: section 2 defines the related terminology, section 3 introduces the smart home kitchen laboratory for algorithm demonstration, section 4 shows the learning of activities with the duration data in a smart home environment, section 5 describes the evaluation framework, section 6 reports the findings of the experiments and discussion based on the results, section 7 discusses related work and section 8 concludes the paper with duration for future work.

## 2 Terminology

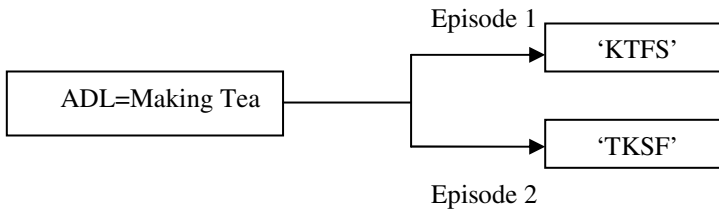
An activity can be considered as an *ADL* completed by a *Person* using a sequence of steps, i.e. an *Episode*, at a given *Time* and for a particular *Duration*. This set of attributes is an extension of previous work in which *Person*, *ADL*, *Episode* and *Time* were used to model the activity pattern [12]. The details of each attribute are described as follows.



**Person:** We consider the possibility of multi-inhabitants in the home. For example, more than one patient may be staying in the same house or family members or a caregiver may be staying with the patient. To track such cases each person is given a unique identity so that they can be distinguished from others during the training phase.

**ADL:** ADLs are defined as the activities carried out in daily living such as eating, bathing, dressing, grooming, etc [13]. There is a set of key ADLs that can be monitored to analyse a person’s health deterioration.

**Episode:** An episode is a particular sequence of sensor activations followed to complete an activity, and indicates a specific way of carrying out an activity. This sequence is extracted from the low-level sensor information received as a stream of sensor activations during the activity. The same activity can be represented by different episodes. Figure1 illustrates the scenario of making tea with two different episodes. In the kitchen objects are attached with different sensors. For instance a tilt sensor is attached to the kettle, contact switches to the tea bag container, sugar container and fridge. For simplicity of presentation the initial letter of the object name is taken to represent the sensor attached to it, for example: tea bag container = ‘T’, kettle= ‘K’, fridge=’F’ and sugar container =’S’. The episode ‘KTFS’ thus means, the kettle sensor is activated, then the tea bag container sensor, then the fridge sensor and finally the sugar container sensor.



**Fig. 1.** Making of tea with two different episodes

**Time:** Time denotes the time of the day at which the activity is carried out. Time is categorised as ‘Morning’, ‘Afternoon’ or ‘Evening’ according to [12].

**Duration:** We consider duration as the length of time taken to complete the activity by the inhabitant. Based on the amount of time taken, duration is categorised as ‘Short’, ‘Medium’ or ‘Long’. It can be calculated from sensor time-stamp data, where the door sensor is activated at the start of an activity and again at the end of the activity.

**Definition 1:** A schema *S* of datacube *D* has five dimensions: *Person*, *ADL*, *Episode*, *Time* and *Duration*. Each attribute *j* has a set of domain values  $\{c_1^{(j)}, \dots, c_{k_j}^{(j)}\}$ , where  $k_j$  is the number of unique values for the attribute. Thus each cell *i* of datacube *D* is

represented by  $v_{paetd}$ , which is the cartesian product of five attributes in the form  $\{c_{i_p}^{(Person)} \times c_{i_a}^{(ADL)} \times c_{i_e}^{(Episode)} \times c_{i_t}^{(Time)} \times c_{i_d}^{(Duration)}\}$ .

**Definition 2:** A datacube  $D$  is defined of type schema  $S$  with each cell  $v_{paetd}$  containing the value  $n_{paetd}$ , representing the number of occurrences of the corresponding combination of values from  $Person$ ,  $ADL$ ,  $Episode$ ,  $Time$  and  $Duration$ .

We first present a novel pictorial representation of the five dimensional data. Each cell of the datacube,  $v_{paetd}$ , is the combination of values from  $Person$ ,  $ADL$ ,  $Episode$ ,  $Time$  and  $Duration$ , and is represented by a path. Figure 2 depicts one such possible combination of a cell  $v_{paetd}$ , with attribute value of  $Person=$  'Emma',  $ADL=$  'making coffee',  $Episode=$  'CK', where Coffee='C' and Kettle='K',  $Time=$  'Afternoon' and  $Duration=$  'Short'. This approach can be extended to ' $n$ ' dimensional data, with each axis representing one of the attributes.

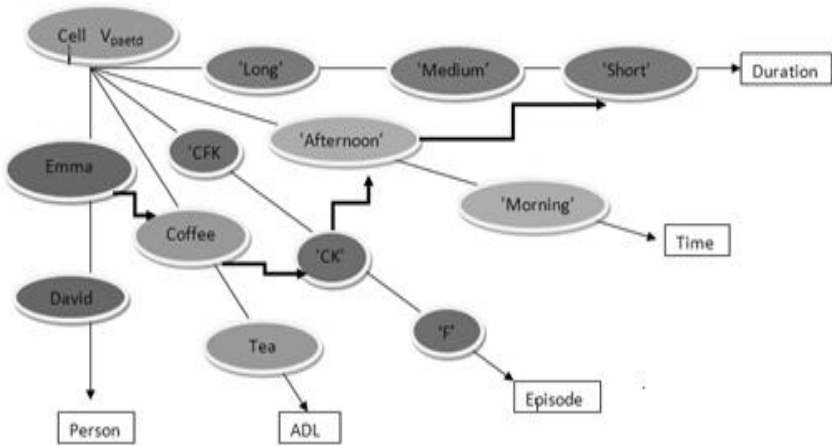


Fig. 2. Shows 'Emma' making 'Coffee' using Episode 'CK' in 'Afternoon' in 'Short' duration

### 3 The Smart Kitchen Laboratory

The experiment was carried out in a smart kitchen laboratory located at the University of Ulster at Jordanstown (Figure 3 (a)). The activity of 'making a drink' was monitored within the environment. The corresponding sensor is triggered when an action is performed. The initial value of the sensor is '0' and changes to '1' when the related object is used. Contact switches were embedded in the related objects (Figure 3(b)) and movement detectors were used to track the presence of an inhabitant. Table 1 shows the list of objects to which sensors are attached along with their type and their representation in the collected data.

**Table 1.** List of sensor

Sensor name	Attached to the object	Type
'C'	Coffee container	Contact switch
'T'	Tea bag container	Contact switch
'S'	Sugar container	Contact switch
'K'	Kettle	Contact switch
'F'	Fridge	Contact switch
'D'	Door	Monitor switch

Real data were collected for a one month period with this set-up. In our experiment, the task of 'making a drink' refers to nine possible activities [20]:

- 1) ADL<sub>1</sub>= 'making black tea'
- 2) ADL<sub>2</sub> 'making tea with milk'
- 3) ADL<sub>3</sub>= 'making tea with sugar',
- 4) ADL<sub>4</sub>= 'making tea with both milk and sugar.'
- 5) ADL<sub>5</sub>= 'making black coffee'
- 6) ADL<sub>6</sub>= 'making coffee with milk'
- 7) ADL<sub>7</sub>= 'making coffee with sugar',
- 8) ADL<sub>8</sub>= 'making coffee with both milk and sugar'
- 9) ADL<sub>9</sub>= 'making a cold drink'

At a higher level the task of 'making a drink' can be categorised as HA1= 'making a cup of tea', HA2= 'making a cup of coffee' and HA3= 'making a cold drink'. For this experiment two people participated. We use virtual names to represent the two users as P1= 'Emma' and P2= 'David'. In this experiment we learn and monitor the behaviour patterns in carrying out the activity. A user interface is designed to record the labelled data. For the process of consistency a set of labelled data is obtained, the user selects labels of *Person* and *ADL* before they start the activity during the training phase. A 'start' and 'end' button is used in order to record the timestamp of an activity and this information is used to partition the sequence of sensor activation into episodes and also capture the activity duration. When sensors are activated, time-stamped data are sent to the server and stored in the database.



**Fig. 3.** The smart kitchen laboratory in the University of Ulster at Jordanstown: (a) an overall kitchen view; (b) a coffee jar attached with a contact sensor

We carry out a discretisation process in order to divide the continuous interval of duration into discrete values. There are several discretisation methods such as visual inspection, equal-width discretisation, cluster-based discretisation, etc. more of this can be found at [14]. In our experiment we follow a simple approach to discretise duration data based on visual inspection of the distribution of duration values [21] and we identify three classes, ‘Short’, ‘Medium’ and ‘Long’, based on clusters within the distribution (Figure 4). Table 2 shows an example of data collected.



**Fig. 4.** The distribution of duration values from collected dataset and discretisation into three categories separated by two vertical lines

**Table 2.** Example data collected

<i>Activity Id</i>	<i>Person</i>	<i>ADL</i>	<i>Episode</i>	<i>Time</i>	<i>Duration</i>
19	David	‘making tea with sugar’	‘TFKS’	Morning	Medium
24	Emma	‘making black coffee’	‘CK’	Afternoon	Short

## 4 Learning Activities of Daily Living in a Smart Home Incorporating Duration

The work is carried out in three stages: firstly, the data collection and pre-processing stage in which the low-level sensor data and the labelled data are integrated to provide data for learning; secondly, the training stage to build the inhabitants’ activity pattern model; thirdly, the evaluation stage, in which the derived model is tested with the new series of simulated data. At the training stage, patterns of inhabitants carrying out activities of daily living are learned. Data labels are required in order to inform the model about inhabitant identities and activities for corresponding data on episode, time and duration. The learned activity model is a joint probability distribution over the different activities represented by the cells in the datacube  $D$ .

After learning the model, recognition can be carried out to predict the person and the activity, based on the learned model and current observation of sensor data, which contains information about the episode, time and duration. A probabilistic learning approach is proposed to obtain inhabitants’ behavioural patterns defined by the joint probability distribution over typical activities in various contexts. Given the data collected in a smart home environment over a period of time, the distribution is obtained using maximum likelihood estimation. The parameter is denoted by  $\pi_{paetd}$  as the probability of cell  $v_{paetd}$  in data cube  $D$ , representing the ‘*Person=p, ADL=a, Episode=e, Time = t and Duration =d*’.  $n_{paetd}$  is the corresponding cardinality for the

number of occurrences of this attribute values combination. Since the aggregates in the datacube  $D$  follow a multinomial distribution, the likelihood is given by equation (1):

$$L \propto \prod_{p=1}^P \prod_{a=1}^A \prod_{e=1}^E \prod_{t=1}^T \prod_{d=1}^D \pi_{paetd}^{n_{paetd}} \quad (1)$$

The maximisation of this likelihood is obtained by setting  $\frac{\partial L}{\partial \pi_{paetd}} = 0$  Subject to the constraint:

$$\sum_{p=1}^P \sum_{a=1}^A \sum_{e=1}^E \sum_{t=1}^T \sum_{d=1}^D \pi_{paetd} = 1 \quad (2)$$

This gives the model; the parameter estimate is as follows:

$$\pi_{paetd} = \frac{n_{paetd}}{N}, \text{ where } N = \sum_{p=1}^P \sum_{a=1}^A \sum_{e=1}^E \sum_{t=1}^T \sum_{d=1}^D n_{paetd} \quad (3)$$

## 5 Evaluation Framework

### 5.1 Performance Criterion

In order to monitor the wellbeing of the inhabitants it is necessary that the model predicts the person and the activity accurately. Therefore the evaluation criteria should be the accuracy of activity recognition from the low level sensor information and the derived model from the training stage.

The prediction of class (*Person, ADL*), (*PA* for abbreviation) indicates what activity is performed and by whom. The derived model is in the form of a probability distribution over activities with their class information (i.e. probability of combination of person, ADL, episode, time and duration stored in each cell). Therefore, the activity prediction is carried out using Equation (5) for the given observation of data, consisting of sensor sequence  $e^o$ , timestamp  $t^o$  and duration  $d^o$ .

$$Pr(p_i, a_j | e = e^o, t = t^o, d = d^o) = \frac{Pr(p_i, a_j, e^o, t^o, d^o)}{\sum_{e=1}^E \sum_{t=1}^T \sum_{d=1}^D Pr(p, a, e^o, t^o, d^o)} \quad (4)$$

This can be solved to give:

$$Pr(p_i, a_j | e^o, t^o, d^o) = \frac{\pi_{p_i a_j e^o t^o d^o}}{\sum_{p=1}^P \sum_{a=1}^A \pi_{p, a, e^o t^o d^o}} \quad (5)$$

where,  $\pi_{p_i a_j e^o t^o d^o}$  is the probability of the person  $p_i$  carrying out activity  $a_j$  at time  $t^o$  via episode  $e^o$ , with duration  $d^o$  of the cell  $v_{p_i a_j e^o t^o d^o}$ . The denominator  $\sum_{p=1}^P \sum_{a=1}^A \pi_{p, a, e^o t^o d^o}$  is the sum of all the possible combinations of *Person* and *ADL* values, for a given episode  $e^o$ , timestamp  $t^o$ , and duration  $d^o$ .

The prediction is then assigned to the class with the highest probability of *Person* and *ADL*:

$$(P, A) = \underset{p_i, a_j}{\arg \max} Pr(p_i, a_j, e^o, t^o, d^o) \quad (6)$$

Classification performance is evaluated by the prediction accuracy, defined as the number of observations for which both the activity and the individual who carried out the activity are correctly identified, in relation to the total number of activity observations in the evaluation dataset.

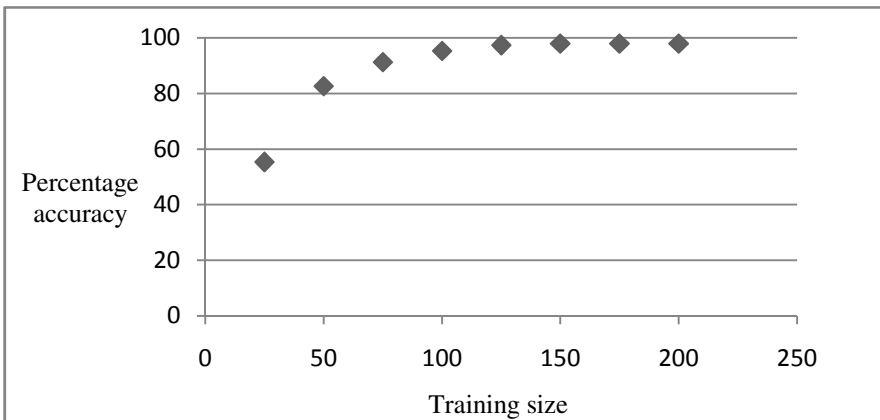
## 5.2 Data Simulation

For the purpose of evaluation we simulate data based on the patterns observed in the real data collected. Generating synthetic data for the evaluation of learning algorithm permits evaluation in a controlled way. Synthetic data are generated based on inhabitants' behavioural model, parameters of an overall probability distribution and the size of data samples. In a datacube  $D$  of schema  $S$ , each cell  $v_{paetds}$  is the number of random values lying in the corresponding interval based on probability distribution. Data values are generated between 0 and 1 in Matlab [15]. The generated random values are categorised according to the probability distribution. Thus each interval value represents the observation that corresponds to a particular combination of *Person*, *ADL*, *Episode*, *Time* and *Duration*. The categorised random data are then aggregated and the cardinality for each cell is the number of random values that fall into the corresponding interval. Thus the simulated data follows the correct probability distribution. We separately generate synthetic data using different seeds for training and evaluation.

## 6 Evaluation Result and Discussion

### 6.1 Results

Performance is evaluated for models learned from different sizes of training dataset. In each case the result is averaged over 10 repetitions to obtain the mean value and the standard deviation. We tested each of the model generated with 150 test dataset.



**Fig. 5.** Prediction accuracy for different size of training dataset

Prediction accuracies are shown in Figure 5 for different sizes of the training dataset. There is an increase in performance as the training data size is increased as expected. After N=150, the model shows stability and there is no further increase in accuracy, so training dataset of size 150 would be sufficient, but we have conservatively fixed the training dataset size as 175.

The derived model is tested for different sizes of test data sets and the performance is compared with the model learned without the duration information [15]. Figure 6 shows an increase in prediction accuracy when the duration information is included in the model.

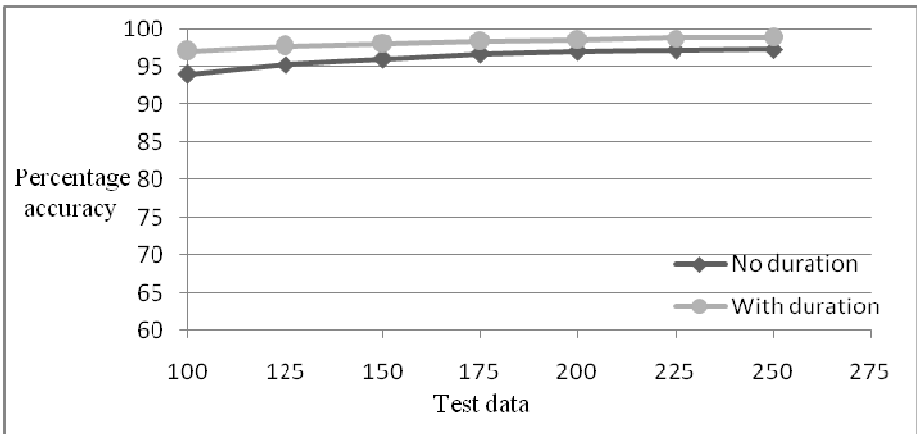


Fig. 6. Model performance for different size of test dataset

## 6.2 Discussion

The derived model gives a better performance when the duration information is incorporated. Including duration information changes the probability distribution and the model is able to distinguish strongly between the ambiguous cases in comparison to the model developed without incorporating duration information. The increase in confidence of activity recognition and person identification is encouraging for developing systems which are more personalised according to individual requirements.

## 7 Related Work

Activity duration provides an additional and informative source of data which is obtained from the sensors. However, exploring this feature for activity recognition is limited in the literature of activity recognition.

Duration information can be used to monitor the wellbeing of the person. For example, if a person usually takes a particular duration to complete an activity but there is an increase in duration for completing the same activity, this may indicate a possible deterioration in the health of the person. The duration of one activity may be

related to another activity that can be represented by Allen's temporal logic. Two or more activities can be overlapping with each other with respect to duration. Allen's temporal logic was used to define thirteen basic relationships between the events based on duration [17]. Temporal rules can be inferred from the duration of the activity. Normal activity can be modelled as temporal checks which can be deployed in a smart home to predict an abnormal condition. If an activity exceeds its predefined duration, it is reported as an abnormality [18]. Time and duration can act as metadata for other sensor data. These temporal semantics can be applied to data so that the same sensor values can be mapped to different activities based on time of day and duration of sensor activations [16]. Duration information can also be used for monitoring the hazardous conditions of the device [19]. If the device is left unattended for long periods, it is reported as a hazardous conditions and an alarm is generated. Duration information is therefore useful for enhancing intelligence and solving complex problems.

## 8 Conclusion and Future Work

The work described here is preliminary but reflects the potential of using duration information in activity prediction. The duration information along with other observed data can help out in recognition. In future, we plan to extend this work in building a reminder system from the partially observed data thus providing assistance to the inhabitant.

## References

1. Pollack, M.E., Brown, L., Colbry, D., McCarthy, C., Orosz, C., Peintner, B., Ramakrishnan, S., Tsamardinos, I.: Autominder: An Intelligent Cognitive Orthotic System for People with Memory Impairment. *Robotics and Autonomous Systems* 44, 273–282 (2003)
2. Relation between severity of Alzheimer's disease and costs of caring.  
<http://www.ncbi.nlm.nih.gov/pmc/articles/PMC1229640/>
3. Mabileau, P., Rahal, Y.: Location Estimation in a Smart Home: System Implementation and Evaluation Using Experimental Data. *International Journal of Telemedicine and Applications* 2008, Article ID 142803, 9 (2008)
4. Cook, D.J.: Health Monitoring and Assistance to Support Aging in Place. *The Journal of Universal Computer Science* 12(2), 15–29 (2006)
5. Giroux, S., Lussier-desrochers, D., Lachappelle, Y.: Pervasive behaviour tracking for cognitive assistance. In: *Proceedings of the 1st ACM International Conference on Pervasive Technologies Related to Assistive Environments, PETR 2008, Athens, Greece, vol. 282(86)* (2008)
6. Martin, T., Majeed, B., Lee, B., Clarke, N.: Group AI. A Third-Generation Telecare System using Fuzzy Ambient Intelligence. *Artificial Intelligence* 175(1), 155–175 (2007)
7. Gao, J., Hauptmann, A.G., Bharucha, A., Wactlar, H.D.: Dining Activity Analysis Using a Hidden Markov Model. In: *17th International Conference on Pattern Recognition (ICPR 2004)*, pp. 2–5 (2004)
8. Pauwels, E.J., Salah, A.A., Tavenard, R.: Sensor Networks for Ambient Intelligence. In: *IEEE 9th Workshop on Multimedia Signal Processing* (2007)



9. Philipose, M., Fishkin, P.K., Perkwowitz, M., Patterson Donald, J., Dieter, F., Henry, K., Inferrin, H.D.: Activities from Interactions with Objects. *IEEE Pervasive Computing Magazine* 3(4), 50–57 (2004)
10. Wu, J., Osuntogun, A., Choudhury, T., Philipose, M., Rehg, J.M.: A Scalable Approach to Activity Recognition based on Object Use Export. In: *IEEE 11th International Conference on Computer Vision, ICCV 2007*, pp. 1–8 (2007)
11. El-zabadani, H., Helal, S., Mann, W., Schmaltz, M., Science, I.: PerVision: An integrated Pervasive Computing/Computer Vision Approach to Tracking Objects in a Self-Sensing Space. In: *Proceedings of the 4th International Conference on Smart Homes and Health Telematic (ICOST)*, Belfast, Northern Ireland, pp. 315–318 (2006)
12. Zhang, S., McClean, S.I., Scotney, B.W., Hong, X., Nugent, C.D., Mulvenna, M.D.: Decision Support for Alzheimer’s Patients in Smart Homes, in *CBMS*, pp. 236–241. *IEEE Computer Society, Jyväskylä* (2008)
13. Definition of ADLs,  
<http://www.medterms.com/script/main/art.asp?articlekey=2152>
14. Yang, Y., Webb Geoffrey, I., Wu, X.: Discretization Methods. In: *Data Mining and Knowledge Discovery Handbook*, pp. 113–130. Springer, US (2005)
15. Matlab®, the language of Technical Computing (Service Pack 3), in Version 7.1.0.246 (R14). The MathWorks, Inc. (2005)
16. Ye, J., Clear, A.K., Coyle, L., Dobson, S.: On using temporal features to create more accurate human-activity classifiers. In: *20th Conference on Artificial Intelligence and Cognitive Science*, pp. 274–283. UCD, Dublin (2009)
17. Jakkula, V.R., Cook, D.J.: Using Temporal Relations in Smart Home Data for Activity Prediction. In: *International Conference on Machine Learning Workshop on the Induction of Process Models (IPM/ICML2007)*, Corvalis (June 2007),  
<http://wwwkramer.in.tum.de/ipm07/> (Cited November 2009)
18. Jakkula, V.R., Cook, D.J.: Learning temporal relations in smart home data. In: *Proceedings of the Second International Conference on Technology and Aging, Canada* (2007)
19. Moncrieff, S., Venkatesh, S., West, G., Greenhill, S.: Multi-modal emotive computing in a smart house environment. *Pervasive and Mobile Computing* 3(2), 74–94 (2007)
20. Hong, X., Nugent, C., Mulvenna, M., McClean, S., Scotney, B., Devlin, S.: Evidential fusion of sensor data for activity recognition in smart homes. In: *Pervasive and Mobile Computing*, pp. 236–252 (2008) (in press)
21. Zhang, S., McClean, S.I., Scotney, B.W., Chaurasia, P., Nugent, C.D.: Using duration to learn activities of daily living in a smart home environment smart home environment. In: *4th International ICST Conference on Pervasive Computing Technologies for Healthcare, IEEE Xplore Digital Library, Munich* (2010)

# A Graphical Model for Risk Analysis and Management

Xun Wang and Mary-Anne Williams

Innovation and Enterprise Research Laboratory  
Centre for Quantum Computing and Intelligent Systems  
University of Technology, Sydney  
{xuwang,mary-anne}@it.uts.edu.au

**Abstract.** Risk analysis and management are important capabilities in intelligent information and knowledge systems. We present a new approach using directed graph based models for risk analysis and management. Our modelling approach is inspired by and builds on the two level approach of the Transferable Belief Model. The *credal* level for risk analysis and model construction uses beliefs in causal inference relations among the variables within a domain and a *pignistic*(betting) level for the decision making. The risk model at the credal level can be transformed into a probabilistic model through a *pignistic transformation* function. This paper focuses on model construction at the credal level. Our modelling approach captures expert knowledge in a formal and iterative fashion based on the Open World Assumption(OWA) in contrast to Bayesian Network based approaches for managing uncertainty associated with risks which assume all the domain knowledge and data have been captured before hand. As a result, our approach does not require complete knowledges and is well suited for modelling risk in dynamic changing environments where information and knowledge is gathered over time as decisions need to be taken. Its performance is related to the quality of the knowledge at hand at any given time.

## 1 Introduction

Risk is inherent in almost every aspect of life, and as a result, risks need to be taken into consideration when information and knowledge systems make decisions. Being able to deal with risks rationally and effectively is crucially important for an intelligent system to perform its functions and achieve its objectives. In relation to dealing with risks, there is a considerable body of work already on causal inference[1] and decision making[2] under uncertainty based mainly on Bayesian Networks (BN) which relies on probability, statistical inference[3]. Construction of a BN based model requires the availability of abundant and accurate data in an environment where every relevant variable is known. Many domains and scenarios simple do not meet these requirements for building a risk model in this way. Capturing existing human knowledge of the domain and environment becomes critical in designing and developing models for risk analysis and management. The BN based approaches do not provide clear and formal development

methodologies for the risk model construction for these cases. Specifically, there is no clear and formalised ways to formulate the required subjective probabilities from inputs of human experts and deal with possible missing domain knowledge (i.e. ignorance). We introduce a formal approach of iterative construction of a graphical causal model for risk modelling and analysis based on the Transferable Belief Model(TBM)[4]. Our risk model is first built at the credal level where we capture the beliefs of the causal relations among the variables associated with risks in the domain with inputs from human experts. We can carry out our analysis and model construction process without being overly concerned with model normalisation and missing knowledge. In addition, our model construction process can accommodate and fuse conflicting inputs from different sources. We also introduce a ranked structure and a rank combination operator to rate the *causal significance* of each causal inference captured in our model. This ranked structure is used to filter out insignificant causal relations when we generate the final graphical model. Our modelling approach supports iterative refinement; it can modify existing risk models using the same method as the initial model construction. We can, therefore, easily port and reuse existing models for similar domains or environments instead of rebuilding the models from scratch. When it is necessary to make decisions, we will transform the risk model to a probabilistic model, i.e. normalise our model, at the lower pignistic level using a pignistic transformation function<sup>1</sup>.

In the following section, we will first introduce a general definition of *risk* and the associated concept of *scenario* for intelligent system. We introduce a simple benchmark problem of ball passing between two soccer playing robots and provide a detailed analysis and description of the problem from a risk management perspective. The graphical model introduced in the later sections is specifically adapted to adhere to this risk perspective. In section 3, we will give a brief background description of the TBM to highlight key concepts and ideas used in our approach. We will illustrate our approach to risk model construction, step by step, using this benchmark problem, highlighting several interesting features of our model. Finally, we give a short discussion on our approach by comparing it with Bayesian network based approaches. We show that our modelling approach is not only consistent and complementary to the BN based approaches, but that it offers significant advantages because it supports iterative construction of the causal inference model. Our plans for future work will also be discussed.

## 2 Risk Modelling

### 2.1 Definition

The concept of risk has not been well defined for general applications. Although people share a general notion of risk, it often carries different technical meanings in different domains and can be interpreted from different perspectives. In order to develop a risk analysis and modelling process for risk management for

---

<sup>1</sup> We will give detailed discussion in subsequent paper.

intelligent information systems that supports or enacts decisions, we develop a practical definition of risk by combining common and essential notions of risk. Our risk model construction process is based upon this definition as given below:

**Definition 1.** A *risk* is a combination of the uncertainty of occurrence of a possible outcome from an initial event and the associated positive or negative payoff<sup>2</sup> of the outcome on our intelligent agent with respect to achieving its goal(s).

**Definition 2.** A *scenario* is the possible outcome or event associated with a risk.

In accordance with the definition, we focus on the two key properties of risk namely, uncertainty and consequences associated with possible scenarios. Both properties are strongly dependent on the task domain, system capabilities and the environment. We use the following benchmark problem to further our discussions.

## 2.2 Benchmark Problem – Autonomous Robot Ball Passing

RoboCup<sup>3</sup> has been one of the driving forces behind advancing and applying theoretical ideas in AI to real world applications and pushing the boundaries of AI. Robot soccer can be viewed as, essentially, a risk management problem, even though it is rarely described in those terms. One of the major challenges in robot soccer matches is ball passing between two robot teammates because it involves making decisions that require the ability to reason about risk. There is still little deliberate ball passing between robots after many years of competitions. The few successful ball passing examples are usually unintentional. This ball passing problem presents a rich scenario in which for example, it is practically impossible to measure your opponent strength; the environment and rule for soccer match constantly evolves. This enables the exploration and analysis of various risk factors and events involved in passing a ball and building risk models of increasing sophistication. It is also an excellent benchmark risk management problem because it is a real world problem where empirical data and experimental results can be collected and the performance of risk modelling methods can be examined, compared, tested, and evaluated.

Our risk modelling process is a stepwise process: domain analysis is conducted first, so that a ranked knowledge base is constructed. This knowledge base can be revised in the same spirit as belief revision, and it can be used to create a graphical belief model, from which a probabilistic model may be generated for final decision making.

---

<sup>2</sup> In risk analysis and management literatures, ‘consequence’ is widely used instead of ‘payoff’. We use ‘consequence’ with the word ‘payoff’ interchangeably.

<sup>3</sup> [www.robocup.org](http://www.robocup.org)

### 2.3 Risk Domain Analysis

We first analyse and describe the problem in terms of *Task/Goal*, the *Environment*, the *Initial Event*, all possible *Scenarios* and all other *Associated Factors*. This simple domain analysis and problem description technique is significantly influenced by Aven[5].

**Task/Goal:** Passing a ball between two NAO robots<sup>4</sup>.

**Environment:** A RoboCup NAO soccer match with two opposing teams and each team is comprised of several identical robots.

**Initial Event:** One robot attempts to kick a ball towards one of its teammate, the receiver.

**Scenarios:** Final possible outcomes of the initial event are summarised in the Table 1. They are simplified scenarios, which allow us to highlight important features of our risk modelling approach.

**Table 1.** A simplified analysis of possible scenarios

SCENARIO	DESCRIPTION
$S_1$	Ball kicked and caught by $R_B$
$S_2$	Ball kicked and intercepted by an opposition robot.
$S_3$	Failed to kick the ball.

**Associated Factors/Variables:**

- Distance ( $D$ ): Distance between robot  $R_A$  and robot  $R_B$  is 20 centimetres (in our example instance).
- Nearby Robots ( $NR$ ): Any nearby robots (either friendly or hostile excluding  $R_B$ ) could possibly intercept the ball.

Concepts described under these five categories form the basic building blocks of our risk model. Further analysis and additional information helps us link these concept variables together and form an ever increasingly complete model as new information is acquired. The risk model for ball passing presented here is a simplified version of the model that we would use in a real soccer match environment. It is sufficiently complex to illustrate the main ideas and our risk model construction process.

## 3 Theoretical Foundation

### 3.1 Assumptions

Assume we have a finite propositional language  $\mathcal{L}$  with the usual representations of tautology and contradictions, and is closed under usual Boolean connectives. We adopt the possible world approach such that our beliefs held for the domain

<sup>4</sup> <http://www.aldebaran-robotics.com/en>

are true in a set of worlds. Let  $\Omega$  be a set of worlds that corresponds to the interpretation of  $\mathcal{L}$ . We further assume that our concepts described under the five categories with our risk analysis can be *well defined* using our language  $\mathcal{L}$ . That is, there is no partial overlap between these concepts and they are self-consistent. In other words,  $\Omega$  is well *partitioned* for our task domain. We will let  $A$ s be the atoms that form the partitions of  $\Omega$ . Our language should also be sufficient to describe all possible relationships between the concepts.

### 3.2 Transferable Belief Model

Development of our risk model requires a way to represent the uncertainties associated with the *scenarios* and *associated factors* developed in earlier analysis. TBM is a model for representing the quantified beliefs for uncertainty. Beliefs could be quantified credibility, subjective support, strength of opinions etc. Beliefs are represented with probability functions or belief functions. TBM consists of two representation levels: the credal level where beliefs are used for modelling and reasoning, and the pignistic level where beliefs are considered and used to make decisions[6]. Since the use of probability functions is only required when a decision or betting is involved, in TBM, probability functions are used at the pignistic level for making decisions, whereas belief functions are used to quantify uncertainty and to reason at the credal level. One important advantage of using belief functions at the credal level is that normalisation of beliefs is less of a concern. Translation between the two levels is achieved through the use of pignistic transformation functions[7]. In this paper, our focus is on building a causal model at the credal level. The following sections give a brief overview of the key concepts in TBM that will play important roles in our modelling.

**Belief Function:** Belief functions is used to model uncertain knowledge. From the possible world perspective: our  $\Omega$  is the frame of discernment, that is, a finite set of mutually exclusive elements. Let  $\mathcal{R}$  be the power set of  $\Omega$  which consists of  $2^\Omega$  of  $\Omega$ . A belief function is a function  $bel$  maps  $\mathcal{R}$  to  $[0,1]$  such that:

1.  $bel(\emptyset) = 0$ ;
2.  $bel(A_1 \cup A_2 \cup \dots \cup A_n) \geq \sum_i bel(A_i) - \sum_{i>j} bel(A_i \cap A_j) \dots - (-1)^n bel(A_1 \cap A_2 \cap \dots \cap A_n)$ , for all  $A_1, A_2, \dots, A_n \in \mathcal{R}$ .

One of the key concepts in TBM is the *basic belief mass* (bbm) where the total amount of specific support for  $A$  is provided by the sum of all bbm given to subsets of  $A$ . When relevant new information comes in, the bbm that initially allocates to  $A$  may be *transferred* to the subsets of  $A$ . This redistribution of support is called *specialisation*. As in TBM, the Dempster's rule of conditioning and combination rule are cases of specialisation[6]. Based on the definition of belief function and basic belief mass, we can now introduce several key concepts found in TBM. These key concepts also form the basis for our model.

**Simple Support Function:** A simplified version of a belief function is called a simple support function. It consists of only two non-null basic belief masses.

One gives support to  $\Omega$ ; the other gives support to a specific subset of  $\Omega$ ,  $A$ . We let  $A^x$  be the simple support function with belief mass of  $m(A) = 1 - x$  and  $m(\Omega) = x$ . That is, we have belief support  $1 - x$  for proposition  $A$ ; and  $x$  support for all other possible worlds that do not include  $A$ .

**Vacuous Belief Function:** In TBM, we can represent total ignorance with a vacuous belief function. A vacuous belief function is a belief function that  $m(\Omega) = 1$ , hence,  $bel(A) = 0 \forall A \in \mathcal{R}, A \neq \Omega$  and  $bel(\Omega) = 1$ . This means, belief in all possible worlds and there is no specific belief for any subset of worlds. The vacuous belief function is usually denoted as  $T$ .

**Latent Belief Structure:** Another key feature in TBM is its ability to represent the conflicting belief of “there are some reasons to believe  $A$ ” and “there are some reasons *not* to believe  $A$ ” simultaneously. TBM uses a latent belief structure to represent such a state of belief. A latent belief consists of a pair of belief functions  $(C, D)$  where  $C$  and  $D$  represent a confidence and a diffidence component of the structure respectively. The classical belief state can be represented by  $(C, T)$  where  $T$  is the vacuous belief function. It has only the confidence component. Whereas,  $(T, D)$  has only the diffidence component which means we only have reasons *not* to believe  $A$ . These pure belief states can also be represented using simple support functions such as  $(A^x, T)$  and  $(T, A^y)$ , where  $x, y \in [0, 1]$  and are compliments of the weights corresponding to “some reasons”.

**The  $\Lambda$  Operator and Apparent Belief:** We need to ensure that, as the overall effect, the confidence component and diffidence component can effectively counter-balance each other; when they are numerically equal, they cancel each other out completely. A  $\Lambda$  operator[4] was introduced to transform a latent belief structure into an apparent belief function such that  $\Lambda(A^x, A^x) = T$ . The apparent belief function is effectively a normal belief function. In the later section, we propose a modification of the  $\Lambda$  operator:  $\hat{\Lambda}$  which computes the relative strength of a causal inference.

## 4 Dynamic Risk Model Construction and Revision

The primary visual representation for our risk model is a directed graph grouped under an initial event such as Fig. 1(a), with nodes representing concept variables<sup>5</sup> in the application domain, and arcs between the nodes representing the inference relations between the variables which capture the notion of causality. Nodes and causal inference relations between the nodes are stored in a knowledge repository as sentences. We employ a ranked structure in this knowledge repository to capture the relative order of the causal inferences according to their causal strengths. With the acquisition of new knowledge regarding the relations between nodes, we combine existing and new beliefs, and then revise the rankings of relevant inference; consequently maintaining the consistency of the knowledge

<sup>5</sup> We use word *concept variables* or just *variables* for risk factors and scenario.

base. This ranked structure facilitates the generation of the final graphic causal model for analysis. We will use the ball passing problem to illustrate this process in more details.

#### 4.1 Model Concept Variables

We assume those variables identified during the early phase of risk analysis, namely, risk factors and scenarios are *relevant* for our risk model. If a variable is no longer relevant due to the acquisition of new information, i.e. it is no longer has any significant causal relations with other variables in the domain, it is removed it from the system. We also distinguish a *normal node* to represent a risk factor and a *scenario node* to represent a final outcome (i.e. a leaf node). Graphically, a normal node is represented by a circle whereas a scenario node is represented by a round cornered square. In our problem, we have an initial setup of two robots separated by a Distance(D) of 20 centimetres, with scenarios of  $S1$  and  $S2$ . All these concepts are stored in our knowledge base. Graphically, they are the single value nodes as shown in Fig. 1(a), along with the initial causal inferences linking between them.

#### 4.2 Model Causal Inference

We insist on capturing the causal relationships among the relevant variables within the domain because such a model represents the stable knowledge of the domain. More importantly, one of main goals of risk management is to be able to take proper interventions to minimise the consequences from the undesirable outcomes. This goal is only achievable when we understand the causal relations.

**Lead:** The causal inference relations between nodes carry the uncertainty information in our model. An arc between two nodes is coupled with a basic belief mass  $m$  to represent the associated uncertainty. We call such a causal inference relation a *Lead*, denoted as  $L_{X \rightarrow Y}$ , where the arc starts from node  $X$  and ends at node  $Y$ . Note that we can only have leads between normal nodes and leads from normal nodes to scenario nodes. No leads can initiate from the scenario nodes to normal nodes or other scenario nodes. To ensure we capture causal inference relations, we formally define *Lead* using the Ramsey test[8]. That is,

**Definition 3.** For a knowledge base  $K$ ,  $X$  and  $Y$  are two simple random variables. We accept a **lead**  $L_{X \rightarrow Y}$  if and only if  $Y$  is accepted (with bbm  $m$ ) in  $K * X$ , where  $K * X$  denotes ‘ $K$  revised by  $X$ ’.

This formal definition is illustrated through the following questions: “Based on what you know about a soccer match ( $K$ ), if robot  $R_A$  kicked the soccer ball towards robot  $R_B$  at distance  $D$ , will you accept the belief that the ball will be caught by  $R_B$  in situation  $S1$ ?” and “What value ( $m$ ) do you put on this belief?”. The value you attribute might be related to your confidence in the belief or related beliefs. A lead  $L_{D \rightarrow S1}$  follows immediately with the answers. In fact, these questions can be readily used to capture opinions from domain experts.



**Definition 4.** A *vacuous lead* is a lead  $L_{X \rightarrow Y}$  with  $bbm\ m = 0$  denoted as  $T_{X \rightarrow Y}$ .

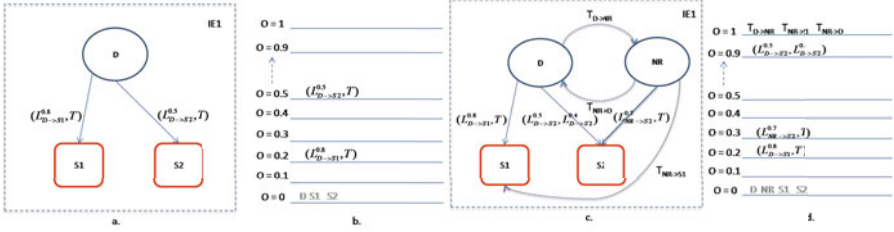
It means we are ignorant of whether there is a causal inference relation from node  $X$  to  $Y$ .

**Frame of Discernment  $\Omega_X$ :** We define a frame of discernment for every variable in our graph model. For a normal node  $X$ , a frame of discernment,  $\Omega_X$ , is a set of all possible leads initiated from  $X$  to other nodes. Since a scenario (leaf) node can have no leads, its frame of discernment is an empty set. Therefore, the frame size for a normal node is  $n - 1 + m$  where  $n$  is number of normal nodes and  $m$  is number of scenario nodes; and the frame size for a scenario node is zero. For our example (as in Fig. 1(a)), the frame of discernment for node  $D$  is  $\Omega_D = \{L_{D \rightarrow S1}, L_{D \rightarrow S2}\}$  in which  $L_{D \rightarrow S1} = 0.8$  and  $L_{D \rightarrow S2} = 0.5$ . It literally means that we have some reasons to believe kicking at distance of 20 centimetres  $D$  leads to scenario  $S1$  with a belief of 0.8.

**Latent Lead:** We can go one step further and employ a structure similar to the latent belief structure in TBM so that we have an additional diffidence component in a latent lead. With this diffidence component we can represent the cases that we have some reasons **not** to believe node  $X$  leads to  $Y$ . For example,  $(L_{D \rightarrow S1}^{0.8}, L_{D \rightarrow S1}^{0.5})$  means we have reason to believe (with  $bbm\ 0.8$ )  $D$  will lead to scenario  $S1$ ; and at the same time, we also have reasons to believe (with  $bbm\ 0.5$ ) our  $D$  will **not** lead to scenario  $S1$ . This is particularly useful, since incoming new information can reinforce or weaken or even dismiss the existing support for the leads between the nodes. With this latent lead structure, we can handle conflicting inputs from different knowledge sources. We also have a more powerful way to express the causal relationships between variables such as expressing “negative” probabilities<sup>6</sup> for the causal relations which is not possible under normal BN like probabilistic models. In addition, the latent lead structure provides a compact form for storing the inference relations in our knowledge base. Latent lead structure  $(L_{D \rightarrow S1}, T_{D \rightarrow S1})$  and  $(T_{D \rightarrow S1}, L_{D \rightarrow S1})$  represent the two special cases in which we have full confidence with zero diffidence and no confidence with full diffidence respectively. They are semantically equivalent to  $X \vdash Y$  and  $X \not\vdash Y$  respectively in the knowledge base. When the confidence component and diffidence component cancel each other out, for example if we have inputs from two experts with exactly opposite views (with equal weighting) on  $NR$  would lead to  $S1$  such that  $(L_{NR \rightarrow S1}^{0.2}, L_{NR \rightarrow S1}^{0.2})$ , then we are left with a vacuous lead  $T_{NR \rightarrow S1}$ . That is, we are totally ignorant whether there should be any causal inference from  $NR$  to  $S1$ . Note, a vacuous lead is implicitly present in a graph normally and it should not be drawn to avoid cluttering of arcs. As in our Fig. 1(a), we emphasis their existence by represented them as dashed arcs.

---

<sup>6</sup> Intuitively, this could be regarded as *resistance* towards having the causal inference relation.



**Fig. 1.** Initial graph setup (a) and its corresponding ranked structure (b). Evolution of the graphic model (c) and its corresponding ranked structure (d).

**Ranked structure:** One of the key features of our modelling approach is using a ranked structure to store causal inference relations according to their relative causal strengths Fig. 1(b). Our knowledge repository (theoretically) captures all possible causal relations (of various strengths) among all variable nodes. We set up the ranking system from 0 to a maximum rank of 1. Rank 0 is given to those sentences representing the causal inference relationship that are *definitely* plausible to our task domain. Sentences that are the least causally plausible with respect to our domain, i.e. vacuous leads should always have the rank of 1. They represent the relations that we are totally ignorant of in our domain. All causal relations of which we do not have specific information are assumed to be vacuous leads. With this ranked structure, we have a clear picture of relative strengths of causal relationships between various risk factors and scenarios in our model. More importantly, we can filter out weak leads when we conduct reasoning and decision making under limited computational power.

**The  $\hat{A}$  Operator:** Accompanying the ranked structure, we also introduce a transform operator  $\hat{A}$  to map latent lead structures into rank values (i.e. causal strength) similar to the  $A$  operator in TBM. The  $\hat{A}$  operator have following properties:

$$0 \leq \hat{A}((L_{x \rightarrow y}^u, L_{x \rightarrow y}^v)) \leq 1 \tag{1}$$

$$\hat{A}((L_{x \rightarrow y}^u, L_{x \rightarrow y}^v)) = 1 \text{ if } u = v. \tag{2}$$

We may have different  $\hat{A}$  operators designed for different domain environments. For example, we could put different discounting factors on confidence and diffidence components. As for our ball passing example, a simple transform operator may be used:

$$\hat{A}((L_{x \rightarrow y}^u, L_{x \rightarrow y}^v)) = 1 - |v - u| \tag{3}$$

where  $u, v$  are in range of  $[0,1]$ .

### 4.3 Construction of an Initial Risk Model

With all the necessary elements described in the previous sections, we can derive a simple algorithm for construction of an initial risk model. This construction

algorithm is slightly modified for later model revisions. Note that, it is not necessary to add vacuous leads into the knowledge base in practice. They are implicitly assumed when we have the relevant variables in the knowledge base. One of the key features of our modelling approach is easy revisions on the existing risk model as new information is acquired. Revisions of the risk model are of two kinds: revision of variables in the model and revision of causal inference relations of the model. Both revision operation can result structural change in the graphic model.

---

**Algorithm 1.** Initial Risk Model Construction Algorithm

---

**Require:** Analysis and describe the problem domain under the five categories (as in section 2.3). Relevant variable nodes form a set  $\mathcal{V}$ .

- 1: **for all**  $N \in \mathcal{V}$  **do**
- 2:   Add the node  $N$  to the knowledge base.
- 3:   **if**  $N$  is a normal node **then**
- 4:     **for all** possible  $L_{N \rightarrow X} \in \Omega_N$  **do**
- 5:      Add possible (vacuous) leads  $T_{N \rightarrow X}$  (virtual step).
- 6:      Solicit input from the knowledge source based on the definition of lead (i.e. the Ramsey test).
- 7:      **if** we have  $L_{N \rightarrow X}$  with bbm  $m \neq 0$  **then**
- 8:       Fuse  $L_{N \rightarrow X}$  into the knowledge base at the ranking calculated using a combination rule and  $\hat{A}$  operator.
- 9:      **end if**
- 10:    **end for**
- 11:    **else if**  $N$  is a scenario node **then**
- 12:     do nothing.
- 13:    **end if**
- 14: **end for**
- 15: Prune all vacuous leads and nodes connected only with vacuous leads (optional step).

---

**4.4 Revision of Concept Variables**

Adding or removing a variable node can occur when the domain(or environment) of the risk model changes. Such a change triggers reanalysis of the domain. Existing risk factors or scenarios may be removed and additional risk factors or scenario may be added. Algorithm 2 for node addition, in fact, is a slightly modified version of algorithm 1. Addition of variables implicitly adds all possible vacuous leads between existing nodes and the newly added node. Fig. 1(c) shows, for example, when node  $NR$  is added to an initial setup of Fig. 1(a), we automatically add four vacuous leads of  $T_{D \rightarrow NR}$ ,  $T_{NR \rightarrow D}$ ,  $T_{NR \rightarrow S1}$  and  $T_{NR \rightarrow S2}$  which represent new possibilities. Intuitively, it makes sense that when we get to know the notion of Nearby Robots for the first time, we have no idea how  $NR$  is related to the existing nodes of  $D$ ,  $S1$  and  $S2$ . As new pieces of evidence for these leads are acquired, they may become valuable leads and their ranks can move below 1. Removing a node means removing all leads between

the retiring node and rest of nodes (algorithm 3). Another important property of variable revision is that addition or removal a node does not affect the rest of leads that are not associated with the node. That is, no additional computation is required to adjust these existing leads. In particular, removal of a node means that bbms associated with any leads from other nodes into the node are transferred to their respective  $m(\Omega)$ ; other leads in the frame are not disturbed. In Fig. 1(a), removal of  $S2$  means  $L_{D \rightarrow S2}$  is also removed. The bbm associated with this lead is transferred to  $m(\Omega_D)$  and the only remaining lead  $L_{D \rightarrow S1}$  in  $\Omega_D$  is unchanged. This is only possible due the minimum commitment principle used in TBM.

---

**Algorithm 2.** Node Addition Algorithm
 

---

**Require:** An existing risk model. Adding variable node  $N$ .  $\mathcal{V}_n$  is a set of all existing normal nodes.

- 1: Add the node  $N$  to the knowledge base.
  - 2: **if**  $N$  is a normal node **then**
  - 3:   **for all** possible  $L_{N \rightarrow X} \in \Omega_N$  **do**
  - 4:     Solicit input from the knowledge source based on the definition of lead (i.e. the Ramsey test).
  - 5:     **if** we have  $L_{N \rightarrow X}$  with bbm  $m \neq 0$  **then**
  - 6:       Fuse  $L_{N \rightarrow X}$  into the knowledge base at the ranking calculated using a combination rule and  $\dot{A}$  operator.
  - 7:     **end if**
  - 8:   **end for**
  - 9: **end if**
  - 10: **for all**  $X$  in  $\mathcal{V}_n$  **do**
  - 11:   Solicit new input on lead  $L_{X \rightarrow N}$ .
  - 12:   **if** we have  $L_{X \rightarrow N}$  with bbm  $m \neq 0$  **then**
  - 13:     Fuse  $L_{N \rightarrow X}$  into the knowledge base at the ranking calculated using a combination rule and  $\dot{A}$  operator.
  - 14:   **end if**
  - 15: **end for**
- 

#### 4.5 Revision of Causal Inferences

Revision of causal inference relations simply means fusing the existing information of a lead with new information about the inference relation and then shuffling the lead within the ranked structure. For example, after addition of node  $NR$ , new information that gives support of 0.7 to the lead from  $NR$  to  $S2$  is combined with the existing vacuous lead  $T_{NR \rightarrow S2}$  using the Dempster's rule of combination such that:

$$\begin{aligned}
 & (L_{NR \rightarrow S2}^{0.7}, T_{NR \rightarrow S2}) \oplus T_{NR \rightarrow S2} \\
 &= (L_{NR \rightarrow S2}^{0.7}, T_{NR \rightarrow S2}) \oplus (T_{NR \rightarrow S2}, T_{NR \rightarrow S2}) \\
 &= (L_{NR \rightarrow S2}^{0.7}, T_{NR \rightarrow S2})
 \end{aligned}$$

---

**Algorithm 3.** Node Removal Algorithm

---

**Require:** An existing risk model. Removing variable node  $N$ .  $\mathcal{V}_n$  is a set of all existing normal nodes.

- 1: **for** each lead  $L$  in the knowledge base **do**
  - 2:   **if**  $L \in \Omega_N$  **or**  $L \in \{L_{X \rightarrow N}, \forall X \in \mathcal{V}_n, X \neq N\}$  **then**
  - 3:     Remove  $L$ .
  - 4:   **end if**
  - 5: **end for**
- 

Similarly, if we have another incoming negative support, say 0.4, for lead  $L_{D \rightarrow S2}$ , i.e.  $(T_{D \rightarrow S2}, L_{D \rightarrow S2}^{0.4})$ , applying the combination rule, we have the result latent lead as  $(L_{D \rightarrow S2}^{0.5}, L_{D \rightarrow S2}^{0.4})$ . Using  $\dot{A}$  operator of Eqn. 3, we compute the new rankings for the latent leads and make appropriate rank adjustments within the ranked structure. Consequently, the ranked structure evolves from Fig. 1(b) to Fig. 1(d). We can visualise this adjustment process as the invisible arc from  $NR$  to  $S2$  becoming visible while the lead moves lower in the ranked structure; and the arc from  $D$  to  $S2$  “fades away” when the lead moves up in the ranked structure. The same process is repeated after the arrival of any further information.

#### 4.6 Graphical Model Generation

The model manipulation process described above ensures our knowledge repository always remain in a consistent state at the credal level. With the ranking system, we can generate our graphic model selectively with causal inferences of significant strengths. We can use a rank cut-off point to exclude all weak leads from our model to avoid any unnecessary complication<sup>7</sup>. For example, by excluding all vacuous leads, we end up with a graph in Fig. 1(c) *without* the dashed arcs. This filtering process effectively generates graphic models which are approximations of actual knowledge repository. Depending on our attitude towards low probable risks, we could choose a different rank cut-off point for graph model generation depending on the desired risk profile. Formally, generation of a graph involves combining all relevant variable nodes and associated leads (with ranks below the cut-off point) using the TBM disjunctive combination rule[4]. Noted, the frames of discernment in the model are mutually exclusive, hence, the frame of discernment for the graph is set of all possible leads among the normal and scenario nodes. That is,

$$\Omega_{graph} = \bigcup \Omega_X, \forall X \in \mathcal{V}. \tag{4}$$

The frame size  $|\Omega_{graph}| = |\mathcal{V}_n(2^{|\mathcal{V}_n-1|} + \mathcal{V}_s)|$ , where  $\mathcal{V}_n$  is a set of the normal nodes,  $\mathcal{V}_s$  is a set of the scenario nodes and  $\mathcal{V} = \mathcal{V}_n \cup \mathcal{V}_s$ .

---

<sup>7</sup> This is similar to how we humans concentrate on “what is important” when dealing with a complex problem.

## 5 Discussion

In this paper, we develop an intuitive and effective approach and concrete algorithms for generating graphic models for risk analysis and management<sup>8</sup>. Our approach formalises the process of capturing domain expert knowledge from a risk perspective in an iterative fashion. In particular, our technique intends to capture the causal inference relations among the domain variables that are relevant with respect to risks. Most existing approaches for managing uncertainty are probabilistic models. Their constructions and dynamics requires a considerable amount of data as sample inputs[9]. As in our benchmark ball passing problem and many other environments, obtaining abundant and meaningful data is just not feasible. In addition, probabilistic models are inherently based on the Close World Assumption, that is, they are unable to represent ignorances and frequent revision in knowledge in open environments. Our modelling process fills these gaps and formalising a process to capture (potentially conflicting) experts' knowledge and create graphical models for risk reasoning and decision making. It is also sufficiently general to be adapted for analysis and management uncertainty in a wide variety of domains. Furthermore, our TBM based model does not preclude the use of data and probabilities; we can still use probability in place of the belief function in our risk model. Another key feature of our model is that initial model construction and follow-on model modifications all use the same revision mechanism. It is possible to construct a risk model in one particular domain/environment, and revise the model to a similar environment with minimum modifications. Existing probabilistic models evolution are, in essence, model selections[10] based on datasets.

Currently, our approach to model construction and revision is currently restricted to processed information(either by human experts or machines) for variables and leads between variables. However, it is possible to employ data learning techniques developed for BN or machine learning to pre-process raw data. When we use belief mass for ranking evaluation, it is possible to exclude a lead with low belief of occurrence but with a huge potential consequence when we are generating the graphic model. Future work will to develop an appropriate measure for risk, and a richer ranking system to reflect measure of risk. To this end, we need to analyse and study the  $\hat{A}$  operator further and to develop more sophisticated transformation operators. Finally, we plan to implement our model for the soccer ball-passing benchmark problem and to test it empirically, as well as in other possible environments in our future research.

## References

1. Pearl, J.: Causality: Models, Reasoning, and Inference. Cambridge University Press, Cambridge (2000)
2. Oliver, R., Smith, J. (eds.): Influence diagrams, belief nets and decision analysis. Wiley, Chichester (1990)

---

<sup>8</sup> The modelling technique show here works on simple variate nodes, it can be extended for multi-variate variables.

3. Pearl, J.: Probabilistic Reasoning in Intelligent Systems. Morgan Kaufmann Publishers Inc., San Francisco (1988)
4. Smets, P., Kennes, R.: The transferable belief model. *Artificial Intelligence* (66), 191–234 (1994)
5. Aven, T.: Risk Analysis. Wiley, Chichester (2008)
6. Smets, P.: The transferable belief model for quantified belief representation. In: Gabbay, D.M., Smets, P. (eds.) *Handbook of Defeasible Reasoning and Uncertainty Management Systems*, vol. 1, pp. 267–302. Kluwer Academic Publisher, Dordrecht (1998)
7. Smets, P.: Constructing the pignistic probability function in a context of uncertainty. In: Henrion, M., Shachter, R.D., Kanal, L.N., Lemmer, J.F. (eds.) *Uncertainty in Artificial Intelligence*, North Holland, Amsterdam, vol. 5, pp. 29–40 (1990)
8. Lindstrom, S., Rabinowicz, W.: Conditionals and the ramsey test. In: Dubois, D., Prade, H. (eds.) *Handbook of Defeasible Reasoning and Uncertainty Management Systems*, vol. 3, pp. 147–188. Kluwer Academic Publisher, Dordrecht (1998)
9. Zuk, O., Margel, S., Domany, E.: On the number of samples needed to learn the correct structure of a bayesian network. In: *Proceedings of the 22nd Conference on Uncertainty in Artificial Intelligence*. AUAI Press, Arlington (2006)
10. Ramachandran, S.: Theory Refinement of Bayesian Networks with Hidden Variables. PhD thesis, The University of Texas at Austin (1998)

# Towards Awareness Services Usage Characterization: Clustering Sessions in a Knowledge Building Environment

Pedro G. Campos<sup>1,2</sup> and Ruth Cobos<sup>1</sup>

<sup>1</sup> Department of Computer Science, Universidad Autónoma de Madrid, 28049,  
Madrid, España

<sup>2</sup> Department of Information Systems, Universidad del Bío-Bío, Avda. Collao 1202,  
Concepción, Chile

pedro.campos@estudiante.uam.es, ruth.cobos@uam.es

**Abstract.** It is well known that members of a work group need awareness about other members, common elements and group process, and this becomes crucial in computer supported knowledge building systems. This work aims to help in evaluating use and results of a web-based collaborative knowledge building environment's awareness services, using web usage mining techniques, particularly clustering over session logs. Results show that different types of sessions, related with their duration and awareness service usage level, can be identified. These results complemented with survey answers from users show interesting findings about utility and satisfaction levels assigned to different awareness services provided.

## 1 Introduction

The implementation of Knowledge Management, and particularly Knowledge Building systems by itself does not guarantee that knowledge flows efficiently between users. It is required to establish an environment conducive to learning and dissemination of knowledge. Important mechanisms to provide such environments for work groups are awareness services [1]. It is well known that members of a work group need awareness about other members, common elements and group process [2]. In fact, in many real-life situations, the knowledge acquisition and creation processes are encouraged in shared working spaces which allow easy interactions between participants (consider for example the importance of participating in conference meetings, which allow direct, face-to-face interactions).

This is one of the main reasons why the KnowCat system has been recently added with some awareness services for its users. KnowCat, acronym for “Knowledge Catalyser” [3,4], is a fully consolidated and thoroughly tested web-based collaborative knowledge building environment developed at Universidad Autónoma de Madrid (UAM, Spain). This system has been used with several student communities since 1998 with the main aim of generating collaboratively quality collective knowledge. Moreover, some studies carried out with KnowCat users have



given us evidences that these services can facilitate task resolution, and allow users to be aware that they are working in a collaborative way [5]. However, we had neither information about general usage level of these services nor evidence about how these services are being used. Several questions may arise including: what are the main characteristics of KnowCat's user sessions?, which are the most demanded awareness services?, what is the usage level of the awareness services provided?, what factors can influence the usage of these services?

This work aims to help in evaluating use and results of KnowCat's awareness services, using data mining techniques, particularly web usage mining [6], in order to establish reliably the support level given by these services, and ease system administrators decision making about improving or extending services. The rest of this paper is organized as follows: section 2 presents main aspects of awareness services in Computer Supported Collaborative Work (CSCW), and details awareness services provided by the KnowCat system. Section 3 introduces the web usage mining techniques used in the services' usage characterization. Section 4 explains the proposed approach. Section 5 presents a discussion on obtained results, and finally section 6 give some conclusions and future work.

## 2 Awareness Services and KnowCat System

### 2.1 CSCW and Awareness Services

The awareness concept, which comes from the field of psychology, was initially adopted in the context of CSCW. Dourish and Bellotti brings one of the firsts definitions for this concept: "Awareness is an understanding of the activities of others, which provides a context for your own activity. This context is used to ensure that individual contributions are relevant to the group's activity as a whole, and to evaluate individual actions with respect to group goals and progress. The information, then, allows groups to manage the process of collaborative working" [1]. From this definition, awareness can be considered as an artifact for emphasize a sense of belonging to a group in CSCW context. One of the major benefits from awareness services is to facilitate coordination among people, and to provide useful hints for initiating communication and collaboration.

Research in the CSCW field has revealed the importance of groupware applications that are able to integrate efficient group awareness mechanisms. Gutwin and Greenberg consider that group awareness must provide updated information about workspace, participants location and their actions in workspace [7]. Table 1 show awareness elements related with workspace [7].

### 2.2 KnowCat System

KnowCat (Knowledge Catalyzer) is a web-based, collaborative knowledge management system which allows the creation of knowledge areas where relevant, quality knowledge about a specific topic can be found [3,4]. In this environment, interactions and contributions from users are part of the "knowledge crystalization" process in which the knowledge management feature of the system is based

**Table 1.** Workspace awareness elements (adopted from [7])

Elements	Relevant questions
Identity	Who is participating in the activity?
Location	Where are they?
Activity Level	Are they active in the workspace? How fast are they working?
Actions	What are they doing? What are their current activities and tasks?
Intentions	What are they going to do? Where are they going to be?
Changes	What changes are they making? Where changes are being made?
Objects	What objects are they using?
Extents	What can they see ?
Abilities	What can they do?
Expectations	What do they need me to do next?
Sphere of Influence	Where can they have effects?

on. This system has been used as support platform on several subjects given at Universidad Autónoma de Madrid (Spain) and Universitat de Lleida (Spain) [8].

KnowCat provides affordances for collaborative knowledge construction. It encourages communities to share their knowledge and, progressively, construct knowledge sites of reasonable quality. These knowledge sites, accessed through a specific URL, are organised around three knowledge elements: i) a knowledge tree (hierarchical structure of topics); ii) a set of documents contained in each topic, with alternative descriptions of the topic; and iii) a set of annotations contained in each document (comments and opinions about the content document). Users participate in the common task of constructing the community knowledge with a set of operations: a) adding documents; b) voting documents; c) adding annotations to documents; and d) adding a new version of a document (details about KnowCat can be found in [3,4,8]).

### 2.3 Awareness Mechanisms Considered

Recently, KnowCat has been extended with awareness services, which allow users to gather information about how other team members have used the system. These awareness services are shown in the “awareness console” that appears in the bottom part of the KnowCat user interface. They provide to KnowCat’s users the following activity group information:

1. Brief information about *Registered Users* and brief information about *Connected Users*: what have these users done?
2. *Radar View*: where and what are the connected users doing?
3. *Global Participation*: How many times have the registered users done each task?
4. *Fish Eye*: when, where and what has each registered user done?
5. *Note Graph*, a map of interaction among users in the annotating task: who has annotated the document of whom?

### 3 Web Usage Mining

The purpose of Web data mining, also known as Web mining, is the discovering of useful information from the web hyperlink structure, page content and usage data. Although Web mining uses many data mining techniques, it is not purely an application of traditional data mining due to the heterogeneity and semi-structured or unstructured nature of Web data. Web mining task can be categorized into three types: Web structure mining, Web content mining and Web usage mining [9]. This work is based on web usage mining, which corresponds to the discovery of access patterns from usage logs, which register each user petition to the web server. One of the key aspects in Web usage mining is log data preprocessing, needed for producing an adequate input for the mining algorithms.

Using standard data mining techniques such as clustering and association rules, a particular user can be associated to other users who exhibit similar behavioral patterns and preferences. Moreover, web usage mining offers a real validation for web site developers, as they can detect whether user behavior differs from expected behavior, according to the web site design. Web usage mining can be performed at several levels. For example, interest may be on the navigational sequence of a particular user, for personalizing user experience. On the contrary, interest may be on the aggregated behavior of many users (e.g. all users in one week period) for measuring aspects as navigability [10].

Preprocessing is generally the most time-consuming step in data mining. Moreover, this is a primary task in Web usage mining in order to use standard algorithms. The most common preprocessing tasks are [11]: i) data cleaning and filtering; ii) de-spidering; iii) user identification; iv) session identification; and v) path completion.

## 4 Applying Web Usage Mining for Awareness Services Usage Characterization

Prior to results discussion we show the general framework considered for pattern discovery and a brief explanation of preprocessing actions undertaken which highlights some important details needed for obtain patterns allowing answering questions as the ones presented in the introduction.

### 4.1 General Framework

The general framework presented in [11] for Web usage mining is considered in this work. According to it, Web usage mining consist of four phases: i) Input stage, where raw web log files are retrieved as well as registration information (if any) and information concerning the site topology; ii) Preprocessing stage, where raw web logs are transformed into a format conducive to fruitful data

mining; iii) Pattern discovery stage, where statistical and data mining methods are applied for pattern discovery; and iv) Pattern analysis stage, where non-interesting or useful patterns are discarded. Here, human analysts examine the output from pattern discovery stage and glean the most interesting, useful, and actionable patterns.

For step i) the information input correspond to the web server log file, which registers users accesses to KnowCat's available services, including awareness services. For step ii) a series of preprocessing actions were carried out, whose main purpose was to identify user sessions' activities, and formatting the information for data mining algorithms application. Here we consider a user session roughly as the set of web pages viewed (services requested) by a particular user for a particular purpose. Preprocessing actions are described in the next subsection. For step iii) a clustering method was applied, specifically Simple k-Means [12], which is a popular and simple algorithm. For step iv) the researchers analyzed the obtained patterns, adjusting some preprocessing steps and algorithm parameters iteratively, until reaching clear tendencies and patterns which allow to answer questions presented in section 1.

## 4.2 Data Preprocessing

Figure 1 shows a schematic view of steps needed for adequate preprocessing of data. In the data extraction step, all major fields of the log files were separated and inserted in a database table (the http request field is considered just as a text string in this step). The timestamp creation is important to handle time ordering of registers. In the request data extraction step, the request field was parsed in order to obtain important information about petitions made by users, focusing on requested services identification. This can be done given that KnowCat implements services (particularly awareness services) as Perl scripts which must be requested. Table 2 shows KnowCat's awareness services and associated file scripts.

In the first data cleaning step, registers not contributing information (i.e. image petitions) were omitted. The session reconstruction was done by identifying registers with same IP, user platform and browser (as registered in the log), whose accesses were within a time frame of 30 minutes between them. Given that only some registers include information about user identification; this data was propagated to remaining registers within the same session. In order to identify possible concurrent accesses from the same IP (as could happen with users in different computers behind the same proxy, as in university labs), when different users were identified in the same session, all registers in that session were omitted; if none register in a session had information about user identification, those registers were also omitted. This corresponds to second data cleaning step. Path completion was not performed, as most services in KnowCat system are accessible from the main view. The final data in the database table were transferred into a text file in CSV format for data mining algorithms application.

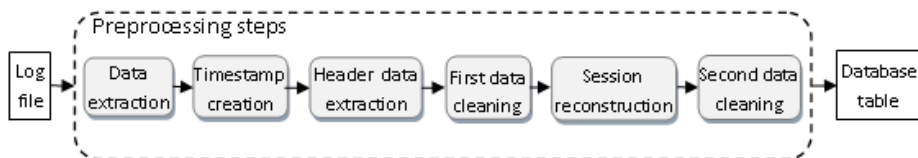


Fig. 1. Preprocessing steps

Table 2. Awareness services and file scripts in KnowCat

Service	File script
Registered users	awConsoleUsrRgs.pl
Connected users	awConsoleUsrCnc.pl
Radar view	awConsoleRdView.pl
Global Participation	awConsoleGlbPr.pl
Note Graph	awConsoleNoteGraph.pl
Fish Eye	awConsoleFishEye.pl

## 5 Results and Discussion

Data was collected from web server log during October, November and December, 2009, corresponding to http requests to KnowCat server (<http://knowcat.ii.uam.es>). We used Weka version 3.6 Simple k-Means algorithm implementation.

### 5.1 Sessions' Characteristics

Table 3 shows an overview of sessions carried out by KnowCat's users during the study time.

It can be seen that the most used awareness service is *Connected users*, followed (far behind) from *Radar View* and *Fish Eye*. The huge petitions value for *Connected users* is mainly due to the fact that this service, when requested,

Table 3. Sessions' Characteristics

Characteristic	Value
Number of sessions (validated)	672
Mean number of actions (requests) per session	253
Mean session duration	2506s. (~41 m.)
Number of sessions in which awareness services were used	650 (96,73%)
<i>Registered users</i> petitions	1166
<i>Connected users</i> petitions	24585
<i>Radar View</i> petitions	2340
<i>Global participation</i> petitions	1986
<i>Note Graph</i> petitions	254
<i>Fish eye</i> petitions	2300

executes a script which generates a register in the log every two minutes until the user logs out from the system. So, it is not an accurate measure about service usage, but it can be considered as a tendency descriptor, as the script is executed only after a user request for that service.

## 5.2 Clustering Sessions

The clustering algorithm, Simple k-Means, was used with  $k = 2, 3$  and 4. The model with  $k = 3$  was found as the most descriptive one, with a sum of squared errors of 32.8 (seed: 10). Table 4 shows the found clusters. Cluster 0 can be described as short duration sessions, in which no awareness service is used. This occurs for a low amount of sessions (3%). An intermediate group (cluster 1), which corresponds to the majority of the sessions (81%) can be described as sessions using awareness services moderately, possibly related to users looking for general information about their classmates whilst doing assigned tasks in Know-Cat. Finally, the third group (cluster 2, 16% of sessions) correspond to sessions with an intensive use of awareness services. These sessions can be described as exploration sessions, where users want to know with detail the work done by their classmates.

**Table 4.** Simple k-Means clustering description

Attribute	Total	Cluster 0	Cluster 1	Cluster 2
Sessions	672	22 (3%)	542 (81%)	108 (16%)
Actions	253.1	12.68	151.09	814.0
Duration	2506.1	193.5	1309.9	8980.2
Awareness services usage	0.967	0	1	1
<i>Registered Users</i> petitions	1.735	0	0.8967	6.2963
<i>Connected Users</i> petitions	36.58	0	15.1	151.8
<i>Radar View</i> petitions	3.482	0	1.0664	16.3
<i>Global participation</i> petitions	2.955	0	1.1993	12.37
<i>Anotation Graph</i> petitions	0.378	0	0.345	0.6204
<i>Fish View</i> petitions	3.42	0	3.1605	5.4352

In order to prevent the analysis of data corresponding to different usage purposes, a course segmentation was done. Table 5 shows results applying Simple k-Means algorithm over data from “Collaborative Systems” course at UAM during Fall 2009 (seed: 7). It must be noted that only this course was using KnowCat as support platform during that time period.

In order to detect temporal differences, we establish date bins for clustering. To obtain representative data, sessions were divided into three time intervals (bins), corresponding each bin to a calendar month in the analyzed time period. The considered intervals are presented in Table 6.

During this period, two main task related with knowledge creation in Know-Cat were assigned to students of the “Collaborative Systems” course. Each of them consisted on adding one document about a given topic (one topic for all

**Table 5.** Simple k-Means clustering description over data of “Collaborative Systems” course

Attribute	Total	Cluster 0	Cluster 1	Cluster 2
Sessions	323	6 (2%)	24 (7%)	293 (91%)
Actions	355.0	3.3333	1983.1	228.9
Duration	3763	190	22011	2341.9
Awareness services usage	0.981	0	1	1
<i>Registered Users</i> petitions	2.025	0	17.125	0.8294
<i>Connected Users</i> petitions	61.36	0	373.75	37.0239
<i>Radar View</i> petitions	4.077	0	40.5	1.1775
<i>Global participation</i> petitions	3.443	0	28.2	1.4846
<i>Anotation Graph</i> petitions	0.146	0	1.1667	0.0648
<i>Fish View</i> petitions	19.4	0	232.6	2.3106

**Table 6.** Date bins considered for clustering

Bin	Date/Time
1	From 10/01/2009-00:00:00 to 10/31/2009-23:59:59
2	From 11/01/2009-00:00:00 to 11/30/2009-23:59:59
3	From 12/01/2009-00:00:00 to 12/31/2009-23:59:59

students, and a different topic in each task); once documents were available in KnowCat, each student had to read the documents of their classmates, making annotations about their content, and voting for the three best documents. Taking into account the annotations, each student had to write a reviewed (second) version of his/her document. Activities related with the first task were performed from october 13 to november 13, 2009. Activities involving the second task were performed from december 3 to 30, 2009. Results of temporal analysis with the Simple k-Means algorithm are presented in Table 7 ( $k = 3$ , seed:40, sum of squared errors=23.16).

**Table 7.** Simple k-Means clustering considering date bins on “Collaborative Systems” course

Attribute	Total	Cluster 0	Cluster 1	Cluster 2
Sessions	323	133 (41%)	99 (31%)	91 (28%)
Actions	355.0	293.6	243.2	566.6
Duration	3763	2871.6	2384.9	6566.6
Date Bin	1.98	1	2	3
Awareness services usage	0.981	0.9699	1	0.978
<i>Registered Users</i> petitions	2.025	2.0902	1.5354	2.4615
<i>Connected Users</i> petitions	61.36	44.5	39.99	109.9
<i>Radar View</i> petitions	4.077	2.7669	3.101	7.0549
<i>Global participation</i> petitions	3.443	3.5338	3.0808	3.7033
<i>Anotation Graph</i> petitions	0.146	0.1429	0.0404	0.2637
<i>Fish View</i> petitions	19.4	3.14	1.7	62.35

Notably, each of the three clusters identified by k-Means corresponded to a different bin (month). It is interesting to observe that all sessions in bin 2 (November) used awareness services, although with low intensity, as in bin 1 (October), whilst a notably increment in the usage of these services can be observed in bin 3 sessions (December). This can be due to 2 main factors: 1) a face-to-face class in which the awareness services available in KnowCat were presented and explained and its usage was encouraged, was performed on november 30, 2009. We think this may be the most influencing factor for the usage increment; and 2) On the last days of December a second task should be deliver on the system, and knowing actions of classmates could result beneficial for the task results. This second factor can explain the notably increment on the *Fish Eye* service usage.

### 5.3 Students' Perception

In order to compare the above results with the students perceptions, we applied a questionnaire about their satisfaction with awareness services. It consisted of 54 questions regarding information about their personal work, work with documents in KnowCat, system usage, learning process and outcomes, group work, course development and awareness services. The questionnaire was applied to 12 students (out of 15) from the "Collaborative Systems" course.

To the question "where have you found instruments facilitating group work?" all students surveyed mention "KnowCat's knowledge space" (KnowCat environment) as first option, and 6 students (50% of the answers) selected the "Awareness console" as second option. To the question "The Awareness console services have helped you to the development of assigned tasks?" 4 students (33%) answer "yes" or "mainly". These answers remark that for some users the awareness services are important tools, although mainly complementary for the majority of the students.

Students were required also to assign a utility level to each service. Interestingly, the *Global Participation* service was found as the most useful (8 students, 67% qualify it as a medium or high utility service). "Registered Users" and "Connected Users" services were also well qualified, with 50% of students assigning them a medium or high utility score. On the other hand, *Radar View*, *Fish Eye* and *Annotation Graph* services were qualified as low or no useful by most students. Similar results were obtained when students were asked to rate the satisfaction with each service. The higher satisfaction was obtained with *Global Participation* service (9 users, 75% rate it as giving them medium or high satisfaction), followed by *Connected Users*, *Registered Users* and *Annotation Graph* services (4 users, 33% rate it as giving them medium or high satisfaction). The last lowest satisfaction level was achieved with *Fish Eye* and *Radar View* services with only 3 and 2 students (25% and 17%) rating them as giving medium or high satisfaction respectively. These findings are somewhat unexpected, as the most requested services according to the log are *Connected Users*, *Radar View* and *Fish Eye* (each of these services totalize more than 2.000 requests, versus less than 2.000 request for each of the remaining services, see Table 3). However,



after analyze the clustering results, It can be seen that the distribution of the request for them are not equally distributed, concentrating on the last month of work (see Table 7) and in the “exploring sessions” cluster (see Table 6); on the contrary, the *Global Participation* service was requested in a more distributed fashion among time and session types. This can be interpreted as that awareness services more attractive to users are used without needing promotion of them, and will not necessarily be the most used, but used permanently. On the other hand, the high number of requests for *Fish Eye* service may be due to some complication in the use of it (the case of *Connected Users* service is not interpreted, as the number of requests obtained is not exact as was stated previously).

## 6 Conclusions and Future Work

This work has presented some results on awareness service usage characterization over a web-based collaborative knowledge building environment. Sessions were clustered in order to obtain information about types of sessions performed and the usage given by students to the awareness services provided by the KnowCat system. Furthermore, a temporal binning of the sessions was made in order to include this dimension in the clustering analysis. The results show a clear distinction between sessions, being possible to identify 3 types of session: i) short-duration, without awareness services usage sessions, possibly dedicated to a brief review of assigned tasks in the systems; ii) medium-duration, with a moderate usage of awareness services sessions (the majority of them), possibly dedicated to task development in the system; and iii) long-duration, with intensive use of awareness services sessions, possibly dedicated to explore and gain information about other users. The temporal analysis also show that encouraging students to use awareness services effectively increments its usage, although a questionnaire applied to students shows that this forced use is not related with a perception of utility or satisfaction. These results allow to answer, at least partially, the questions posed at the beginning of this work.

This information is valuable for many reasons. In first place, it allows to establish if implemented awareness systems are being effectively used, verifying that the effort of its development is recompensed with a proper usage level. Moreover, it can be used to perform some adaptation to the user interface, locating the most used services in preference place, whilst services with very low or no usage can be discarded, leading to a social “crystalisation” of these services.

It is important to note that the tasks carried out can be employed for the evaluation of awareness in other knowledge creation systems as well, with the only requirement of keeping track of the usage of these systems, in a system log for example. The collaborative knowledge building system called Knowledge Practices Environment [13] could take advantage of this evaluation scheme for instance, if the above-mentioned requirement is fulfilled.

However, we must state that this work is just our first attempt to characterize and evaluate the awareness services incorporated to KnowCat. We plan to extend the study including more students in other courses that are supported by the

system, in order to be able to perform more fine-grained evaluation, to a user and task level for example. Furthermore, we consider analyzing other dimensions such as the quality of tasks performed with support of awareness services, or the academical outcomes (grades) of students. The user survey can be improved too, incorporating questions related with the importance of the availability (or not) of each of the awareness services.

Finally, we state that characterizing the effective usage of awareness services in collaborative knowledge building environments is important for several reasons. The main aim of such services is to facilitate communication among users, so the level of utility of services provided should be measured in order to assure this goal. On the other hand, awareness can be implemented in many ways, and new forms of awareness are developed constantly. Thus, knowing the satisfaction of users with implemented services can lead to the renovation of some of them. Also, counting with this kind of assessment tools facilitates experimentation with new forms of awareness, allowing a faster adaptation of state-of-the-art services, or even the development of new user-oriented and system-oriented services.

## Acknowledgements

This research was partly funded by the Spanish National Plan of R+D, project number TIN2008-02081/TIN and by the CAM (Autonomous Community of Madrid) project number S2009/TIC-1650 (CAM). The first author acknowledges support from the Chilean Government (BecasChile scholarship program). We wish to thank the anonymous reviewers for their helpful comments.

## References

1. Dourish, P., Bellotti, V.: Awareness and coordination in shared workspace. In: CSCW 1992: Proceedings of the 1992 ACM Conference on Computer-supported Cooperative Work, pp. 107–114. ACM, New York (1992)
2. Gross, T., Stary, C., Totter, A.: User-centered awareness in computer-supported cooperative work-systems: Structured embedding of findings from social sciences. *Journal of Human-Computer Interaction* 18(3), 323–360 (2005)
3. Alamán, X., Cobos, R.: Knowcat: A web application for knowledge organization. In: ER 1999: Proceedings of the Workshops on Evolution and Change in Data Management, Reverse Engineering in Information Systems, and the World Wide Web and Conceptual Modeling, pp. 348–359. Springer, Heidelberg (1999)
4. Cobos, R.: Mecanismos para la cristalización del conocimiento, una propuesta mediante un sistema de trabajo colaborativo (mechanisms for the crystallisation of knowledge, a proposal using a collaborative system) (2003)
5. Cobos, R., Claros-Gómez, I.D., Moreno-Llorena, J.: A proposal of awareness services for the construction of quality community knowledge supported by the knowledge management system knowcat. In: Proceedings of the Symposium on Human Interface 2009 on Conference Universal Access in Human-Computer Interaction. Part I, pp. 365–374. Springer, Heidelberg (2009)

6. Nasraoui, O., Soliman, M., Saka, E., Badia, A., Germain, R.: A web usage mining framework for mining evolving user profiles in dynamic web sites. *IEEE Trans. on Knowl. and Data Eng.* 20(2), 202–215 (2008)
7. Gutwin, C., Greenberg, S.: A descriptive framework of workspace awareness for real-time groupware. *Comput. Supported Coop. Work* 11(3), 411–446 (2002)
8. Pifarré, M., Cobos, R.: Evaluation of the development of metacognitive knowledge supported by the knowcat system. *Int. Journal on Educational Technology Research and Development* 57(6), 787–799 (2009)
9. Liu, B.: *Web Data Mining: Exploring Hyperlinks, Contents, and Usage Data (Data-Centric Systems and Applications)*. Springer, New York (2006)
10. Markov, Z., Larose, D.T.: *Data Mining the Web: Uncovering Patterns in Web Content, Structure, and Usage*. John Wiley & Sons, Inc., Hoboken (2007)
11. Srivastava, J., Cooley, R., Deshpande, M., Tan, P.N.: Web usage mining: Discovery and applications of usage patterns from web data. *SIGKDD Explor. Newsl.* 1(2), 12–23 (2000)
12. MacQueen, J.B.: Some methods for classification and analysis of multivariate observations. In: Cam, L.M.L., Neyman, J. (eds.) *Proc. of the Fifth Berkeley Symposium on Mathematical Statistics and Probability*, vol. 1, pp. 281–297. University of California Press, Berkeley (1967)
13. Minna, L., Sami, P., Kari, K., Hanni, M., Merja, B., Hannu, M.: Main functionalities of the knowledge practices environment (kpe) affording knowledge creation practices in education. In: *CSCL 2009: Proceedings of the 9th international conference on Computer supported collaborative learning*, International Society of the Learning Sciences, pp. 297–306 (2009)

# Adjusting Class Association Rules from Global and Local Perspectives Based on Evolutionary Computation

Guangfei Yang<sup>1</sup>, Jiangning Wu<sup>1</sup>, Shingo Mabu<sup>2</sup>,  
Kaoru Shimada<sup>2</sup>, and Kotaro Hirasawa<sup>2</sup>

<sup>1</sup> Institute of Systems Engineering, Dalian University of Technology, China

<sup>2</sup> Graduate School of Information, Production and Systems, Waseda University, Japan

**Abstract.** In this paper, we propose an evolutionary method to adjust class association rules from both global and local perspectives. We discover an interesting phenomena that the classification performance could be improved if we import some prior-knowledge, in the form of equations, to re-rank the association rules. We make use of Genetic Network Programming to automatically search the prior-knowledge. In addition to rank the rules globally, we also develop a feedback mechanism to adjust the rules locally, by giving some rewards to good rules and penalties to bad ones. The experimental results on UCI datasets show that the proposed method could improve the classification accuracies effectively.

**Keywords:** Association Rule, Classification, Genetic Network Programming.

## 1 Introduction

In recent years, data mining becomes more and more important to discover knowledge from the huge amounts of data gathered by the modern society. The association rule mining is an important field of data mining [1]. Association rules have been successfully applied to classification, and it could achieve promising accuracy [2] [3] [4]. In the existing methods, the association rules are usually ranked by their support and confidence values. Take CMAR [3] for example, where the association rules are ranked by the following policy: given two rules  $r_i$  and  $r_j$ ,  $r_i$  is ranked higher than  $r_j$  if

1. the confidence of  $r_i$  is greater than that of  $r_j$ , or
2. their confidences are the same, but the support of  $r_i$  is greater than that of  $r_j$ , or
3. both the confidences and supports of  $r_i$  and  $r_j$  are the same, but  $r_i$  has less items in the antecedent part than  $r_j$ .

From intuition, we think that the above ranking policy may be a little naive to explore the potential ability of associative classification. In order to verify such intuition, we could do some initial experiments by designing a new ranking policy and importing some prior-knowledge to re-rank the rules, where the priori-knowledge consists of ranking equations which combine the support and confidence values by various functions. The new ranking policy is: given two rules  $r_i$  and  $r_j$ , calculate the ranking values for both  $r_i$  and  $r_j$  by an equation, and  $r_i$  is ranked higher than  $r_j$  if

1. the ranking value of  $r_i$  is greater than that of  $r_j$ , or
2. the ranking values are the same, but the confidence of  $r_i$  is greater than that of  $r_j$ , or

3. both the ranking values and confidences are the same, but the support of  $r_i$  is greater than that of  $r_j$ , or
4. the ranking values, confidences and supports of  $r_i$  and  $r_j$  are all the same, but  $r_i$  has less items in the antecedent part than  $r_j$ .

By doing some initial experiments on UCI data sets, we find that it is possible to improve the classification accuracy by the new ranking policy. Then, the issue is how to generate the prior-knowledge, i.e., the ranking equations. We make use of an evolutionary method, Genetic Network Programming [5], to automatically generate the proper equations for ranking the rules.

In this paper, we also design a feedback mechanism to improve the classification performance further. The ranking equations could adjust all the rules simultaneously, but this adjustment does not pay attention to each single rule. As a result, some rules are ranked properly, while some others may not be ranked well. The reason behind this problem is that each rule has its own property/feature, and it will be better if we could adjust each single rule based on its unique property. The feedback mechanism is proposed to solve this problem by giving some rewards to the *good* rules and penalizing the *bad* ones. Whether a rule is *good* or *bad* is decided by its actual performance in classification, and since we build our method based on evolutionary computation, we could classify data and check the quality of each rule during every generation of the evolution. As the evolution goes on, the *good* rules will be enhanced gradually to have more power to affect the classification results, while the *bad* rules will be weakened. The equations are generated to adjust the ranking of rules from a **global** point of view, and the feedback mechanism refine the classification power of each rule so that this rule will have a more proper impact on classification from a **local** point of view. The global adjustment together with local refinement could be more powerful, which will be validated by the experimental results.

## 2 Backgrounds

### 2.1 Associative Classification

The concept of association rule was first introduced by Agrawal et al. for market basket analysis [1]. Since its first proposal, it has attracted various kinds of research, and associative classification integrated association rules and classification. The first associative classification method is CBA, proposed by [2]. After that, there are some more extended and improved versions, including emerging pattern approach [6], multiple association rules approach(CMAR) [3], predictive associative rules approach(CPAR) [4], graph classification by frequent subgraph [7], top-k covering rules [8], optimal rules approach [9], instance centric rules generation approach [10]. CBA adopts Apriori [11] to mine a large number of association rules which satisfy user-specified minimum support and confidence thresholds [2]. When classifying the records, each record is only classified by one rule. CMAR [3] differs from CBA in terms of not only the association rule mining method FP-growth [12] but also the multiple rules classification method, that is, one record is classified by several rules, and the experiments show that this approach could achieve better accuracy.

We will try to improve the classic CMAR and here we briefly introduce its main ideas [3]. After the rules are ranked and pruned to build the classifier, when classifying a record, the rules matching this record are found. If all these rules have the same consequent class, then this record is classified with this class. Usually the classes of these rules are different, and the rules are divided into several groups according to their class labels, where there is only one group for one class label and the rules with the same class labels are put into the same group. Each group is assigned with a power by combing the power of each rule in this group. Given rule  $r$  and the training data, suppose the number of records containing the antecedent part of rule  $r$  is  $sa$ , the number of records containing the consequent part of rule  $r$  (i.e., the class label) is  $sc$ , the total number of records is  $n$ , and the power of rule  $r$  is measured by weighted chi-squared value  $w\chi^2(r)$ :

$$w\chi^2(r) = (\min(sa, sc) - \frac{sa \times sc}{n})^2 \times n \times \tau, \quad (1)$$

where,

$$\tau = \frac{1}{sa \times sc} + \frac{1}{sa \times (n - sc)} + \frac{1}{sc \times (n - sa)} + \frac{1}{(n - sa) \times (n - sc)}$$

The power of each group  $g$  is  $w\chi^2(g)$ :

$$w\chi^2(g) = \sum_{r \in g} \frac{\chi^2(r) \times \chi^2(r)}{w\chi^2(r)}, \quad (2)$$

For each record, we find out the rule group with the largest  $w\chi^2$ , and classify the record by this group.

## 2.2 Genetic Network Programming

Genetic Network Programming (GNP) is an extension of GP, which uses directed graphs as genes for evolutionary computation [5]. GNP evolves the graph structure with a pre-determined number of nodes, and it could be quite compact and efficient and never cause the bloat [5]. Here we describe the main ideas of GNP with its application to generate equations [13] [14].

The basic GNP structure consists of three kinds of nodes: Start Node, Judgement Node, and Processing Node. The Start Node indicates where the transition starts, the Judgement Nodes decide the transition directions, and the Processing Nodes generate the ranking equations. Besides three kinds of nodes, we also design a memory structure which is associated with each Processing Node. The memory structure is made up of four parts: Weight Memory, Measure Memory, Operator Memory and Expression Memory. The values in the Weight Memory are real numbers between 0 and 1, which are parameters in equations. The Measure Memory contains a set of measures  $\{sup, conf\}$ , and Operator Memory contains the elements of  $\{+, -, \times, \div, sq, sr, max, min, abs\}$  which is used to combine the measures, where  $sq$  is the abbreviation of square,  $sr$  the

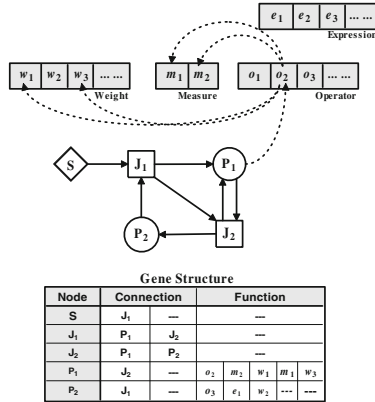


Fig. 1. An example of evolving equations by GNP

square root, *max* the maximum, *min* the minimum, and *abs* the absolute value. Once an equation is generated, it is stored into the Expression Memory. Moreover, the equations in Expression Memory will be used as building blocks to generate more complex equations. For example, once an equation  $(0.97 * sup)/(0.63 * conf)$  has been generated, it may be used to generate another equation  $max\{(0.75 * conf), (0.42 * (0.97 * sup)/(0.63 * conf))\}$ .

After a Processing node is executed, the next node will be a Judgement Node. Each Processing Node is associated with several functional components in memory, such as an operator in Operator Memory, one or two weights in Weight Memory, one or two measures in Measure Memory, and one or two expressions in Expression Memory. After executing a Judgement Node, there will be several possible nodes (either Judgement Nodes or Processing Nodes), and one of them will be activated.

Here is an example in Fig. 1. There are one Start Node (*S*), two Judgement Nodes (*J*<sub>1</sub> and *J*<sub>2</sub>) and two Processing Nodes (*P*<sub>1</sub> and *P*<sub>2</sub>). After *S* is activated, *J*<sub>1</sub> will be the next active node. After *J*<sub>1</sub> is executed, there are two possible nodes, *P*<sub>1</sub> and *J*<sub>2</sub>, and one of them will be the next one. In our simulations, we select each possible node with equal probability for probabilistic transition. *P*<sub>1</sub> is associated with an operator *o*<sub>2</sub>, together with two weight values *w*<sub>1</sub> and *w*<sub>3</sub>, two measures *m*<sub>2</sub> and *m*<sub>1</sub>. After *P*<sub>1</sub> is executed, a new expression, say, *e*<sub>3</sub>, is generated as:  $e_3 = (w_1 \times m_2) o_2 (w_3 \times m_1)$ , where  $w_1, w_3 \in W$ ,  $o_2 \in O$  and  $m_1, m_2 \in M$ .

### 3 Global Ranking and Local Refinement

#### 3.1 Rank Rules by Equations Globally

The flowchart of the proposed method is shown in Fig. 2. Once an equation is determined, all the rules will be ranked by this equation, and we describe it as a *global* approach. Let's explain Fig. 2 with more details.

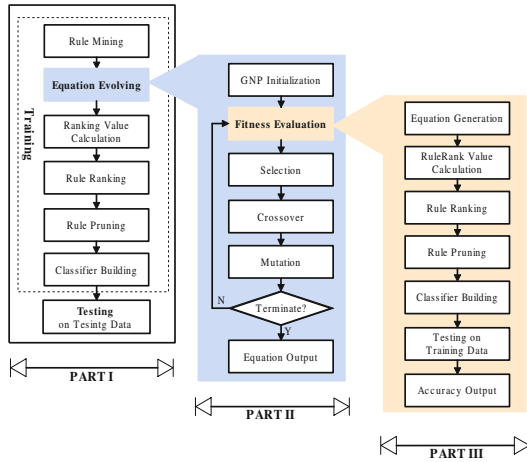


Fig. 2. Flowchart

PART I consists of the most general steps. After the association rules have been mined, we start the evolution to search various kinds of ranking equations, and select the best one to rank rules. After the rules have been ranked, traditional rule pruning and classifier building procedures will be carried out. The details of the second step (i.e., Equation Evolving) will be explained in PART II. The rules are ranked by the new policy introduced before.

PART II shows the steps of the equation evolving procedure, where the technical details of evolution have been introduced before. When a certain number of generations have been reached, the evolution stops and the equation with the best performance will be saved. Each individual usually could generate more than one equation, therefore, we let the best equation to decide the fitness value of the individual. The details of the second step (i.e., Fitness Evaluation) will be explained in PART III.

PART III explains how to evaluate the fitness value of the individuals or calculate the accuracy of the equations. Each equation generated by the evolutionary individuals will be used to rank the rules, and then the rules will be pruned by conventional procedures to perform classification on the training data. Each equation will have an accuracy value accordingly.

### 3.2 Refine Rules by Feedback Mechanism Locally

The ranking equation is to adjust the ranking of association rules from a global point of view. Because the rules are adjusted simultaneously once an equation is determined, it could not pay attention to a certain rule. As a result, maybe some rules are adjusted very well but some others are not. In this section we describe a feedback mechanism to adjust the classification power of each single rule from a local point of view.

*Definition 1.* Given rule  $r$  and record  $d$ , if all the items in the antecedent part of rule  $r$  appear in record  $d$ , then we say  $r$  matches  $d$ , represented as:  $match(r, d) = true$ .



Given a classifier containing a set of rules  $R$ , a data set  $D$ , rule  $r \in R$  and record  $d \in D$ , where  $match(r, d) = true$ , rule  $r$  must belong to one of the following five situations after  $d$  has been classified by  $R$  when applying the multiple rule policy from CMAR:

1.  $R$  has correctly classified  $d$ , while  $r$  has correctly classified  $d$ ;
2.  $R$  has correctly classified  $d$ , while  $r$  has wrongly classified  $d$ ;
3.  $R$  has wrongly classified  $d$ , while  $r$  has correctly classified  $d$ ;
4.  $R$  has wrongly classified  $d$ , while  $r$  has wrongly classified  $d$ , and the classification result of  $R$  and that of  $r$  are the same;
5.  $R$  has wrongly classified  $d$ , while  $r$  has wrongly classified  $d$ , and the classification result of  $R$  and that of  $r$  are different.

For rule  $r$ , we count the number of records in the data set  $D$  corresponding to the first situation as  $v_1(r)$  (i.e., the number of records correctly classified by both  $R$  and  $r$ ), the second as  $v_2(r)$ , the third as  $v_3(r)$ , the fourth as  $v_4^+(r)$  and the fifth as  $v_4^-(r)$ , respectively, as shown in Table 1.  $|r|$  is the total number of records in  $D$  that could be matched by  $r$ .

**Table 1.** Five situations of rule feedbacks

		r		
		correct	wrong	sum
R	correct	$v_1$	$v_2$	$v_1 + v_2$
	wrong	$v_3$	$v_4^+ + v_4^-$	$v_3 + v_4^+ + v_4^-$
	sum	$v_1 + v_3$	$v_2 + v_4^+ + v_4^-$	$ r $

Let’s take a closer look at the meaning of Table 1.  $v_1$  represents both  $r$  and  $R$  has classified the records correctly. For each record, there is a sole correct classification result. If both  $r$  and  $R$  have given correct classification judgements for record  $d$ , it means that they have given the same classification for  $d$ . In other words,  $r$  has joined in the correct judgement for  $d$  by  $R$  and contributed to the accuracy. If rule  $r$  has a large  $v_1$ ,  $r$  must have joined many times in the correct judgements by  $R$  and contributed to the accuracy very much.

In a similar way, we could analyze  $v_4^+$ ,  $v_2$ ,  $v_4^-$  and  $v_3$ .

For each record, there are usually more than one wrong classification results when the number of classes in the data set is more than 2. If both  $r$  and  $R$  have given wrong classification judgements for record  $d$ , there are two cases: (1), the wrong judgements by  $r$  and  $R$  are same(corresponding to  $v_4^+$ ); (2), the wrong judgements by  $r$  and  $R$  are different(corresponding to  $v_4^-$ ). In the first case,  $r$  has joined in the wrong judgement for  $d$  by  $R$  and damaged the accuracy; in the second case, although the judgement by  $r$  is wrong, it has not joined in the wrong judgement for  $d$  by  $R$ , which means  $r$  has not damaged the accuracy. As a result, if rule  $r$  has a large  $v_4^+$ , it must have joined many times in the wrong judgements by  $R$ , and we should weaken  $r$  if we want to improve the accuracy. In contrast, if rule  $r$  has a large  $v_4^-$ , it is difficult to decide whether we should strengthen or weaken  $r$ , and the best choice is to do nothing on  $r$  in this case.

Similarly to  $v_4^-$ ,  $v_2$  means that  $r$  has neither contributed to nor damaged the accuracy. If rule  $r$  has a large  $v_2$  or  $v_4^-$ , we will neither strengthen nor weaken it.

$v_3$  tells us that for a record  $d$ ,  $r$  has given a correct judgement, but  $R$  has given a wrong judgement. It indicates that  $r$  is not powerful enough to affect the classification result by  $R$ . If we increase the power of  $r$ , then judgement by  $R$  may become correct. As a result, if rule  $r$  has large  $v_3$ , this rule has a potential positive impact on the classification, and we will strengthen it.

Before analyzing the feedback mechanism, we return to the question mentioned before: why the new ranking policy proposed in this paper effective? In fact, the answer to this question will be clear if we could answer another question: is it possible that the rules with smaller confidence and support give more correct classification judgements? Based on Table 1, let's review the definitions of confidence and support of rule  $r$ :

$$conf(r) = \frac{v_1 + v_3}{|r|}, \tag{3}$$

$$sup(r) = \frac{v_1 + v_3}{N}, \tag{4}$$

where,  $N$  is the number of records in the data set. The definition of confidence is directly proportional to  $v_1 + v_3$ . If rule  $r$  has a large confidence, it means that  $r$  has a large value of  $v_1 + v_3$ . However, it is  $v_1$  and  $v_4^+$  that measure the truly positive and negative classification contributions to  $r$ , respectively. If rule  $r$  has a large  $v_1$ , it is good for classification, and if  $r$  has a large  $v_4^+$ , it is bad for classification. A large confidence could not tell us whether  $r$  has a large  $v_1$  or  $v_4^+$ , so it is difficult to evaluate the real classification ability of  $r$  by its confidence. It is true that a rule with smaller confidence but larger  $v_1$  or small  $v_4^+$  might give more correct classification judgements. In the same way, it is possible that a rule with smaller support might give more correct classifications sometimes.

Now, let's discuss the feedback mechanism.

*Definition 2.* Given a rule  $r$ , the reward of  $r$ ,  $\psi^+(r)$ , is defined as:

$$\psi^+(r) = \frac{v_3(r)}{v_3(r) + v_4^+(r)}. \tag{5}$$

*Definition 3.* Given a rule  $r$ , the penalty of  $r$ ,  $\psi^-(r)$ , is defined as:

$$\psi^-(r) = \frac{v_4^+(r)}{v_3(r) + v_4^+(r)}. \tag{6}$$

*Definition 4.* Given a rule  $r$ , the feedback of  $r$ ,  $\Psi(r)$ , is defined as:

$$\begin{aligned} \Psi(r) &= 1 + (1 - e^{-\psi^+(r)}) - (1 - e^{-\psi^-(r)}) \\ &= 1 - e^{-\psi^+(r)} + e^{-\psi^-(r)} \\ &= 1 - e^{-\frac{v_3(r)}{v_3(r) + v_4^+(r)}} + e^{-\frac{v_4^+(r)}{v_3(r) + v_4^+(r)}}, \end{aligned} \tag{7}$$

where,  $\Psi(r) > 1$  means that rule  $r$  is to be rewarded, and  $0 < \Psi(r) < 1$  means that rule  $r$  is to be punished.

Please note that in Eq. 5 and Eq. 6, we ignored  $v_1$ . If we consider  $v_1$ , the reward and penalty could be revised as:

$$\psi^+(r) = \frac{v_1(r) + v_3(r)}{v_1(r) + v_3(r) + v_4^+(r)}. \tag{8}$$

$$\psi^-(r) = \frac{v_4^+(r)}{v_1(r) + v_3(r) + v_4^+(r)}. \tag{9}$$

From the analyses up till now, Eq. 5 and Eq. 8 will have almost the same functions, and the same for Eq. 6 and Eq. 9. However, due to the experimental results, we found that Eq. 5 and Eq. 6 were better for improving the accuracies, so we chose these two calculation methods in the experiments.

By the feedback, we adjust the weighted chi-squared value of rule  $r$ :

$$w\chi^2(r) \leftarrow w\chi^2(r) \times \Psi(r), \tag{10}$$

where, the definition of weighted chi-squared  $w\chi^2(r)$  is adopted from CMAR [3].

The weighted chi-squared value decides the ability of each rule to affect the classification results. In other words, a rule with larger weighted chi-squared value is more powerful for classification. By adjusting the weighted chi-squared value of each rule with its feedback, we could increase the ability of the good rules (i.e., the rules which could give more correct classification results), and decrease the ability of the bad rules (i.e., the rules which could give more wrong classification results). As a result, it could be expected that there will be more correct classification, and hence the accuracy could be improved.

## 4 Empirical Evaluations

We implemented the algorithms in JAVA and , and there were totally 23 data sets downloaded from UCI [15]. We evaluated all the algorithms by 10-fold cross validations. Each single experiment was repeated 5 times, and the average performances with standard deviations were calculated.

In Table 2, we compared CMAR, the new rule ranking policy(denoted as RANK) and the new rule ranking policy and the feedback mechanism(denoted as RANK+). From the simulation results, we could see that RANK+ gives the best average accuracy and RANK has the second best average accuracy. Since the feedbacks are directly analyzed from the classification results, and the quality of rules are directly associated with the accuracy, it is natural that the RANK+ is more greedy and could increase the accuracy.

We also give some pairwise comparisons in Table 3, which compare the algorithms one by one to see in how many datasets one algorithm could perform better than another algorithm. From Table 3, we could see that RANK+ is still the best one and RANK is the second best one.

**Table 2.** Comparison of classification accuracy on UCI datasets

Dataset	CMAR	RANK	RANK+
<i>adult</i>	80.73	82.82±0.22	83.27±0.13
<i>anneal</i>	90.09	89.96±0.32	90.25±0.50
<i>breast</i>	89.84	93.12±0.52	93.81±0.19
<i>cleve</i>	82.58	81.80±0.74	81.82±1.47
<i>crx</i>	85.80	85.07±0.83	85.09±0.68
<i>diabetes</i>	72.18	72.18±0.00	72.18±0.00
<i>ecoli</i>	78.02	79.06±0.69	79.67±0.93
<i>flare</i>	84.30	85.01±0.32	85.07±0.39
<i>german</i>	72.00	73.10±0.39	73.56±1.01
<i>glass</i>	51.07	51.02±0.99	51.22±0.49
<i>hepatitis</i>	82.17	81.23±0.90	82.30±0.88
<i>horseColic</i>	81.06	81.12±0.13	81.17±0.15
<i>hypno</i>	95.23	96.80±0.31	96.90±0.16
<i>ionosphere</i>	90.62	90.67±0.13	90.69±0.32
<i>iris</i>	93.33	93.33±0.47	93.73±0.37
<i>labor</i>	86.00	88.00±3.16	88.00±1.41
<i>led</i>	73.31	73.35±0.08	73.35±0.04
<i>page</i>	89.99	90.76±0.23	90.81±0.18
<i>pima</i>	73.85	73.50±0.33	73.82±0.31
<i>sonar</i>	69.86	71.06±1.89	71.46±1.08
<i>waveform</i>	76.22	76.62±0.43	77.12±0.38
<i>wine</i>	92.54	92.06±0.50	92.30±0.49
<i>zoo</i>	94.00	94.00±0.00	94.60±0.55
average	81.95	82.42	82.70

**Table 3.** Pairwise comparison (each number means the method in the row has better accuracy than the method in the column in this number of data sets)

	CMAR	RANK	RANK+
CMAR	-	7	5
RANK	14	-	0
RANK+	18	20	-

## 5 Conclusions

In this paper, we have proposed an evolutionary ranking method and a feedback mechanism for class association rule. The global adjustment and local refinement work together to improve the classification accuracy further. The simulation results show that the proposed method improves the classification accuracy effectively.

## Acknowledgment

This paper is partially supported by the National High Technology Research and Development Program of China(No.2008AA04Z107).

## References

1. Agrawal, R., Imielinski, T., Swami, A.: Mining association rules between sets of items in large databases. In: Proc. of the Int'l Conf. on Management of Data, pp. 207–216 (1993)
2. Liu, B., Hsu, W., Ma, Y.: Integrating Classification and Association Rule Mining. In: Proc. of the Int'l Conf. on Knowledge Discovery and Data Mining, pp. 80–86 (1998)
3. Li, W., Han, J., Pei, J.: CMAR: Accurate and Efficient Classification based on Multiple Class-Association Rules. In: Proc. of the IEEE Int'l Conf. on Data Mining, pp. 369–376 (2001)
4. Yin, X., Han, J.: CPAR: Classification based on Predictive Association Rules. In: Proc. of the Third SIAM Int'l Conf. on Data Mining, pp. 331–335 (2001)
5. Mabu, S., Hirasawa, K., Hu, J.: A Graph-Based Evolutionary Algorithm: Genetic Network Programming (GNP) and Its Extension Using Reinforcement Learning. *Evolutionary Computation* 15(3), 369–398 (2007)
6. Li, J., Dong, G., Ramamohanarao, K.: Making Use of the Most Expressive Jumping Emerging Patterns for Classification. In: Proc. of the 2000 Pacific-Asia Conf. on Knowledge Discovery and Data Mining, pp. 220–232 (2000)
7. Deshpande, M., Kuramochi, M., Karypis, G.: Frequent Sub-structure-based Approaches for Classifying Chemical Compounds. In: Proc. of the 2002 Int'l Conf. on Data Mining, pp. 35–42 (2003)
8. Cong, G., Tan, K.L., Tung, A.K.H., Xu, X.: Mining Top-k Covering Rule Groups for Gene Expression Data. In: Proc. of the 2005 Int'l Conf. on Management of Data, pp. 670–681 (2005)
9. Li, J.: On Optimal Rule Discovery. *IEEE Trans. on Knowledge and Data Engineering* 18(4), 460–471 (2006)
10. Wang, J., Karypis, G.: HARMONY: Efficiently Mining the Best Rules for Classification. In: Proc. of the 2005 SIAM Conf. on Data Mining, pp. 205–216 (2005)
11. Agrawal, R., Srikant, R.: Fast Algorithms for Mining Association Rules. In: Proc. of 20th Int'l Conf. on Very Large Data Bases, pp. 487–499 (1994)
12. Han, J., Pei, J., Yin, Y.: Mining Frequent Patterns without Candidate Generation. *SIGMOD Rec.* 29(2), 1–12 (2000)
13. Mabu, S., Hirasawa, K., Matsuya, Y., Hu, J.: Genetic Network Programming for Automatic Program Generation. *J. of Advanced Computational Intelligence and Intelligent Informatics* 9(4), 430–435 (2005)
14. Yang, G., Shimada, K., Mabu, S., Hirasawa, K.: A Nonlinear Model to Rank Association Rules Based on Semantic Similarity And Genetic Network Programming. *IEEJ Trans. on Electrical and Electronic Engineering* 4(2), 248–256 (2009)
15. UC Irvine Machine Learning Repository, <http://archive.ics.uci.edu/ml/>

# Probabilistic Declarative Process Mining

Elena Bellodi, Fabrizio Riguzzi, and Evelina Lamma

ENDIF – Università di Ferrara – Via Saragat, 1 – 44122 Ferrara, Italy  
{elena.bellodi,evelina.lamma,fabrizio.riguzzi}@unife.it

**Abstract.** The management of business processes is receiving much attention, since it can support significant efficiency improvements in organizations. One of the most interesting problems is the representation of process models in a language that allows to perform reasoning on it.

Various knowledge-based languages have been lately developed for such a task and showed to have a high potential due to the advantages of these languages with respect to traditional graph-based notations.

In this work we present an approach for the automatic discovery of knowledge-based process models expressed by means of a probabilistic logic, starting from a set of process execution traces. The approach first uses the DPML (Declarative Process Model Learner) algorithm [16] to extract a set of integrity constraints from a collection of traces. Then, the learned constraints are translated into Markov Logic formulas and the weights of each formula are tuned using the Alchemy system. The resulting theory allows to perform probabilistic classification of traces. We tested the proposed approach on a real database of university students' careers. The experiments show that the combination of DPML and Alchemy achieves better results than DPML alone.

**Keywords:** Business Process Management, Knowledge-based Process Models, Process Mining, Statistical Relational Learning.

## 1 Introduction

Organizations usually rely on a number of processes to achieve their mission. These processes are typically complex and involve a large number of people. The performance of the organization critically depends on the quality and accuracy of its processes. Thus the processes form a very important asset of organizations and are a fundamental part of their body of knowledge.

The area of Business Processes Management (see e.g. [13]) is devoted to the study of ways for representing and reasoning with process models. Most approaches use forms of graphs or Petri nets [4]. Recently, however, new modeling languages have started to appear that are more knowledge-based and declarative, in the sense that they express only constraints on process execution rather than encoding them as paths in a graph. DecSerFlow [3], ConDec [2] and SCIFF [7,6] are examples of such languages. In particular, SCIFF adopts first-order logic in order to represent the constraints.

The problem of automatically mining a structured description of a business process directly from real data has been studied by many authors (see e.g.

[5,1,14]). The input data consist of execution traces (or histories) of the process and their collection is performed by information systems which log the activities performed by the users. This problem has been called Process Mining or Workflow Mining.

The works [16,15,8] presented approaches for learning models in DecSerFlow/ConDec and SCIFF.

Starting from them, in this paper we present a knowledge-based system to discover declarative logic-based knowledge in the form of business rules, from a set of traces. Process traces are previously labeled as compliant or not: learning a model from both compliant and non compliant traces is interesting if an organization has two or more sets of process executions and may want to understand in what sense they differ.

Additionally the learned process model is able to encode probabilistic information. In fact, the complexity and uncertainty of real world domains require both the use of first-order logic and the use of probability. Recently, various languages have been proposed in the field of Statistical Relational Learning that combine the two. One of these is Markov Logic [19,12], that extends first-order logic by attaching weights to formulas.

We propose to represent process models by means of Markov Logic. Moreover, we present an approach for inducing these descriptions that involves first learning a logical theory with Declarative Process Model Learner [16] and then attaching weights to the formulas by means of the Alchemy system [19].

The effectiveness of the approach is illustrated by considering the careers of real students at the University of Ferrara. The experiment showed that the combined use of DPML and Alchemy for Process Mining outperforms the use of DPML only.

The paper is organized as follows: we first discuss how we represent execution traces and process models using logic programming. Then we presents the learning technique we have adopted for performing Process Mining. After having evaluated the proposed approach on a real world dataset, we discuss related works and conclude.

## 2 Process Mining

A trace  $t$  is a sequence of events. Each event is described by a number of attributes. The only requirement is that one of the attributes describes the event type. Other attributes may be the executor of the event or event specific information.

An example of a trace is

$$\langle a, b, c \rangle$$

where  $a$ ,  $b$  and  $c$  are events executed in sequence.

A *process model*  $PM$  is a formula in a language for which an interpreter exists that, when applied to a model  $PM$  and a trace  $t$ , returns answer yes if the trace is compliant with the description and false otherwise. In the first case we write  $t \models PM$ , in the second case  $t \not\models PM$ .

A bag of process traces  $L$  is called a *log*. The aim of Process Mining is to infer a process model from a log. Usually, in Process Mining, only compliant traces are used as input to the learning algorithm, see e.g. [5,1,14]. We consider instead the case where we are given both compliant and non compliant traces, since non compliant traces can provide valuable information. This is true in particular in the case under study.

## 2.1 Representing Process Traces and Models with Logic

A process trace can be represented as a logical interpretation (set of ground atoms): each event is modeled with an atom whose predicate is the event type and whose arguments store the attributes of the event. Moreover, the atom contains an extra argument indicating the position in the sequence. For example, the trace:

$\langle a, b, c \rangle$

can be represented with the interpretation

$\{a(1), b(2), c(3)\}$ .

Besides the trace, we may have some general knowledge that is valid for all traces. This information will be called *background knowledge* and we assume that it can be represented as a normal logic program  $B^1$ . The rules of  $B$  allow to complete the information present in a trace  $t$ : rather than simply  $t$ , we now consider  $M(B \cup t)$ , the model of the program  $B \cup t$  according to Clark's completion [10].

The process language we consider is a subset of the SCIFF language, originally defined in [6,7], for specifying and verifying interaction in open agent societies.

A process model in our language is a set of Integrity Constraints (ICs). An IC,  $C$ , is a logical formula of the form

$$\begin{aligned} Body \rightarrow \exists (ConjP_1) \vee \dots \vee \exists (ConjP_n) \\ \vee \vee \neg(ConjN_1) \vee \dots \vee \vee \neg(ConjN_m) \end{aligned} \quad (1)$$

where  $Body$ ,  $ConjP_i$   $i = 1, \dots, n$  and  $ConjN_j$   $j = 1, \dots, m$  are conjunctions of literals built over event predicates or over predicates defined in the background knowledge. In particular  $Body$  is of the form  $b_1 \wedge \dots \wedge b_l$  where the  $b_i$  are literals;  $ConjP_i$  is a formula of the form  $event(attr_1, \dots, attr_r) \wedge d_1 \wedge \dots \wedge d_k$  where  $event()$  is an event predicate and  $d_i$  are literals;  $ConjN_j$  is a formula of the form  $event(attr_1, \dots, attr_r) \wedge d_1 \wedge \dots \wedge d_k$ . The quantifiers in the head apply to all the variables not appearing in the body. The variables of the body are implicitly universally quantified with scope the entire formula.

We will use  $Body(C)$  to indicate  $Body$  and  $Head(C)$  to indicate the formula  $\exists(ConjP_1) \vee \dots \vee \exists(ConjP_n) \vee \vee \neg(ConjN_1) \vee \dots \vee \vee \neg(ConjN_m)$  and call them respectively the *body* and the *head* of  $C$ . We will use  $HeadSet(C)$  to indicate the set  $\{ConjP_1, \dots, ConjP_n, ConjN_1, \dots, ConjN_m\}$ .

$Body(C)$ ,  $ConjP_i$   $i = 1, \dots, n$  and  $ConjN_j$   $j = 1, \dots, m$  will be sometimes interpreted as sets of literals, the intended meaning will be clear from the context.

<sup>1</sup> A normal logic program is a program containing clauses of the form  $H \leftarrow B_1, \dots, B_n$  where  $H$  is an atom and the  $B_i$ s are literals, i.e., atoms or negations of atoms.



All the formulas  $ConjP_j$  in  $Head(C)$  will be called  $P$  disjuncts; all the formulas  $ConjN_j$  in  $Head(C)$  will be called  $N$  disjuncts.

An example of an IC is

$$\begin{aligned}
 & order(bob, camera, T), T < 10 \\
 \rightarrow & \exists T1 (ship(alice, camera, T1), \\
 & bill(alice, bob, 100, T1), T < T1 \\
 & \vee \\
 & \forall T1, V \neg bill(alice, bob, V, T1), T < T1
 \end{aligned} \tag{2}$$

The meaning of the IC (2) is the following: if *bob* has ordered a camera at a time  $T < 10$ , then *alice* must *ship* it and *bill bob* 100\$ at a time  $T1$  later than  $T$  or *alice* must *not bill bob* any expense at a time  $T1$  later than  $T$ .

An IC  $C$  is true in an interpretation  $M(B \cup t)$ , written  $M(B \cup t) \models C$ , if, for every substitution  $\theta$  for which  $Body(C)$  is true in  $M(B \cup t)$ , there exists a disjunct  $\exists(ConjP_i)$  or  $\forall\neg(ConjN_j)$  in  $Head(C)$  that is true in  $M(B \cup t)$ . If  $M(B \cup t) \models C$  we say that the trace  $t$  is *compliant* with  $C$ . A process model  $H$  is true in an interpretation  $M(B \cup t)$  if every IC of  $H$  is true in it and we write  $M(B \cup t) \models H$ . We also say that trace  $t$  is *compliant* with  $H$ .

Similarly to what has been observed in [18] for disjunctive clauses, the truth of an IC in an interpretation  $M(B \cup t)$  can be tested by running the query:

$$? - Body, ConjP_1, \dots, ConjP_n, ConjN_1, \dots, ConjN_m$$

against a Prolog database containing the clauses of  $B$  and the atoms of  $t$  as facts. Here we assume that  $B$  is *range-restricted*, i.e., that all the variables that appear in the head of clauses also appear in the body. If this holds, every answer to a query  $Q$  against  $B \cup t$  completely instantiate  $Q$ , i.e., it produces an element of  $M(B \cup t)$ .

If the  $N$  disjuncts in the head share some variables, then the following query must be issued

$$? - Body, ConjP_1, \dots, ConjP_n, ConjN_1, \dots, ConjN_m$$

that ensures that the  $N$  disjuncts are tested separately without instantiating the variables.

If the query finitely fails, the IC is true in the interpretation. If the query succeeds, the IC is false in the interpretation. Otherwise nothing can be said.

## 2.2 Learning ICs Theories

In this section, we briefly describe the algorithm Declarative Process Model Learner (DPML) proposed in [16].

DPML finds an IC theory solving the learning problem by searching the space of ICs. The space is structured using a generality relation based on the following definition of subsumption.

**Definition 1 (Subsumption).** *An IC  $D$  subsumes an IC  $C$ , written  $D \geq C$ , iff it exists a substitution  $\theta$  for the variables in the body of  $D$  or in the  $N$  disjuncts of  $D$  such that*

- $Body(D)\theta \subseteq Body(C)$  and
- $\forall ConjP(D) \in HeadSet(D), \exists ConjP(C) \in HeadSet(C) : ConjP(C) \subseteq ConjP(D)\theta$  and
- $\forall ConjN(D) \in HeadSet(D), \exists ConjN(C) \in HeadSet(C) : ConjN(D)\theta \subseteq ConjN(C)$

If  $D$  subsumes  $C$ , then  $C$  is more general than  $D$ . For example, let us consider the following clauses:

$$\begin{aligned} C &= \text{accept}(X) \vee \text{refusal}(X) \leftarrow \text{invitation}(X) \\ D &= \text{accept}(X) \vee \text{refusal}(X) \leftarrow \text{true} \\ E &= \text{accept}(X) \leftarrow \text{invitation}(X) \end{aligned}$$

Then  $C$  is more general than  $D$  and  $E$ , while  $D$  and  $E$  are not comparable.

The search space is defined by the *language bias* that consists of a set of IC templates, which define the literals that can be added to clauses. In particular, each template specifies:

- a set of literals  $BS$  allowed in the body,
- a set of disjuncts  $HS$  allowed in the head. For each disjunct, the template specifies:
  - whether it is a  $P$  or an  $N$  disjunct,
  - the set of literals allowed in the disjunct.

The search in the space of ICs is performed from specific to general: given an IC  $D$ , the set of refinements  $\rho(D)$  of  $D$  is a set of ICs that are more general than  $D$ . ICs in  $\rho(D)$  are obtained by adding a literal to the body, by adding a disjunct to the head, by adding a literal to an  $N$  disjunct or by removing a literal from a  $P$  disjunct.

The DPML algorithm solves the following learning problem:

**Given**

- a space of possible process models  $\mathcal{H}$
- a set  $I^+$  of positive traces;
- a set  $I^-$  of negative traces;
- a definite clause background theory  $B$ .

**Find:** A process model  $H \in \mathcal{H}$  such that

- for all  $i^+ \in I^+$ ,  $M(B \cup i^+) \models H$ ;
- for all  $i^- \in I^-$ ,  $M(B \cup i^-) \not\models H$ ;

If  $M(B \cup i) \models C$  we say that IC  $C$  *covers* the trace  $i$  and if  $M(B \cup i) \not\models C$  we say that  $C$  *rules out* the trace  $i$ .

Every IC in the learned theory is seen as a clause that must be true in all the positive traces (compliant traces) and false in some negative traces (non compliant traces). The theory composed of all the ICs must be such that all the ICs are true when considering a compliant trace and at least one IC is false when considering a non compliant one.

```

function DPML( $I^+, I^-, B$ )
  initialize  $H := \emptyset$ 
  do
     $C := \text{FindBestIC}(I^+, I^-, B)$ 
    if  $C \neq \emptyset$  then
      add  $C$  to  $H$ 
      remove from  $I^-$  all interpretations that are false for  $C$ 
  while  $C \neq \emptyset$  and  $I^-$  is not empty
  return  $H$ 

function FindBestIC( $I^+, I^-, B$ )
  initialize  $\text{Beam} := \{false \leftarrow true\}$ 
  initialize  $\text{BestIC} := \emptyset$ 
  while  $\text{Beam}$  is not empty do
    initialize  $\text{NewBeam} := \emptyset$ 
    for each IC  $C$  in  $\text{Beam}$  do
      for each refinement  $\text{Ref}$  of  $C$  do
        if  $\text{Ref}$  is better than  $\text{BestIC}$  then
           $\text{BestIC} := \text{Ref}$ 
        if  $\text{Ref}$  is not to be pruned then
          add  $\text{Ref}$  to  $\text{NewBeam}$ 
          if size of  $\text{NewBeam} > \text{MaxBeamSize}$  then
            remove worst clause from  $\text{NewBeam}$ 
     $\text{Beam} := \text{NewBeam}$ 
  return  $\text{BestIC}$ 

```

**Fig. 1.** DPML learning algorithm

The DPML algorithm is an adaptation of ICL [11] and consists of two nested loops: a covering loop (function DPML in Figure 1) and a generalization loop (function FindBestIC in Figure 1). In the covering loop negative traces are progressively ruled out and removed from the set  $I^-$ . At each iteration of the loop a new IC  $C$  is added to the theory. Each IC rules out some negative interpretations. The loop ends when  $I^-$  is empty or when no IC is found.

The IC to be added in every iteration of the covering loop is returned by function FindBestIC. It looks for an IC by using beam search with  $p(\ominus|\overline{C})$  as a heuristic function. The search starts from the IC  $false \leftarrow true$  that is the most specific and rules out all the negative traces but also all the positive traces. ICs in the beam are gradually generalized by using the refinement operator.  $\text{MaxBeamSize}$  is a user-defined constant storing the maximum size of the beam.

At the end of the refinement cycle, the best IC found so far is returned.

### 2.3 Probabilistic Integrity Constraints

Markov Logic (ML) [19] is a language that extends first-order logic by attaching weights to formulas. Semantically, weighted formulas are viewed as templates

for constructing Markov networks. In the infinite-weight limit, ML reduces to standard first-order logic.

**Definition 2 (Markov logic network).** *A Markov logic network (MLN)  $L$  is a set of pairs  $(F_i, w_i)$ , where  $F_i$  is a formula in first-order logic and  $w_i$  is a real number. Together with a finite set of constants  $C = \{c_1, c_2, \dots, c_m\}$ , it defines a Markov network  $M_{L,C}$  as follows:*

1.  $M_{L,C}$  contains one binary node for each possible grounding of each atom appearing in  $L$ . The value of the node is 1 if the ground atom is true, and 0 otherwise.
2.  $M_{L,C}$  contains one feature (real-valued function) for each possible grounding of each formula  $F_i$  in  $L$ . The value of this feature is 1 for a possible world if the ground formula is true in the possible world, and 0 otherwise. The weight of the feature associated to  $F_i$  is  $w_i$ .

For example, an MLN containing the formula  $\forall x \text{Smokes}(x) \rightarrow \text{Cancer}(x)$  (smoking causes cancer) applied to the set of constants  $C = \{\text{Anna}, \text{Bob}\}$  yields the features  $\text{Smokes}(\text{Anna}) \rightarrow \text{Cancer}(\text{Anna})$  and  $\text{Smokes}(\text{Bob}) \rightarrow \text{Cancer}(\text{Bob})$ , and a ground Markov network with 4 nodes ( $\text{Smokes}(\text{Anna})$ ,  $\text{Cancer}(\text{Anna})$ ,  $\text{Smokes}(\text{Bob})$ ,  $\text{Cancer}(\text{Bob})$ ).

A possible world  $\mathbf{x}$  is an assignment of truth values to every ground atom. The probability distribution specified by the ground Markov network  $M_{L,C}$  over possible worlds  $\mathbf{x}$  is given by

$$P(\mathbf{x}) = \frac{1}{Z} \exp \left( \sum_{i=1}^F w_i n_i(\mathbf{x}) \right) \quad (3)$$

where  $F$  is the number of formulas in the MLN,  $n_i(\mathbf{x})$  is the number of true groundings of  $F_i$  in  $\mathbf{x}$ ,  $Z$  is a partition function given by  $\sum_{\mathbf{x}} \exp \left( \sum_{i=1}^F w_i n_i(\mathbf{x}) \right)$  that ensures that  $P(\mathbf{x})$  sums to one.

A set of ICs can be seen as a “hard” first-order theory that constrains the set of possible worlds: if a world violates even one formula, it is considered impossible. The basic idea in Markov Logic is to soften these constraints, so that when a world violates one of them it is just less probable, but not impossible. The weight associated to each formula reflects how strong the constraint is: the higher the weight, the greater the difference in probability between a world that satisfies the formula and one that does not, other things being equal.

Once an IC theory has been learned from data, integrity constraints are transformed into ML formulas and weights are learned for them using the discriminative weight learning algorithm of [12] that is implemented in the Alchemy system<sup>2</sup>.

Each IC of the form (1) is translated into the following ML formula:

$$\begin{aligned} & \text{Body} \wedge \neg(\text{Conj}P_1) \wedge \dots \wedge \neg(\text{Conj}P_n) \\ & \wedge (\text{Conj}N_1) \wedge \dots \wedge (\text{Conj}N_m) \rightarrow \text{neg} \end{aligned} \quad (4)$$

<sup>2</sup> <http://alchemy.cs.washington.edu/>

where *neg* means that the trace is negative. In absence of disjuncts in the head, the IC  $Body \rightarrow false$  reduces to  $Body \rightarrow neg$ . The head of all the formulas always contains only the atom *neg*, while all disjuncts in the head are moved to the body.

An example of IC referred to the analyzed domain is:

$$\begin{aligned} & true \\ \rightarrow & \forall A \neg registration(A, 2005) \\ & \vee \\ & \forall B, C \neg enrollment2(B, C, oc). \end{aligned}$$

This IC states that the students who graduated (positive traces) do not present registration in the year 2005 or an enrollment in the second year as an out-of-course student.

The translation into a formula in Markov logic is:

$$registration(A, 2005) \wedge enrollment2(B, C, oc) \rightarrow neg$$

After weight learning the formula results:

$$1.08555 \quad registration(A, 2005) \wedge enrollment2(B, C, oc) \rightarrow neg$$

with a real number (the weight) attached to the body. The resulting MLN, composed of a set of such formulas, can then be used to infer the probability of *neg*, that is the probability that the trace is negative, given a database consisting of atoms representing the trace.

### 3 Experiments

Our goal is to demonstrate that the combined use of DPML, for learning an IC theory, and Alchemy, for learning weights for formulas, produces better results than the sharp classification realized by the IC theory alone.

The experiments have been performed over a real dataset regarding university students, where the careers of students that graduated are positive traces and the careers of students who did not finish their studies are negative ones. We want to predict whether a student graduates on the basis of her career. To perform our experiments, we collected 813 careers of students enrolled at the Faculty of Engineering of the University of Ferrara from 2004 to 2009. The traces have been labeled as compliant or non compliant with respect to the classification specified above. There are 327 positive and 486 negative traces.

We first induce an IC theory from these data. Every trace was therefore adapted to the format required by the DPML algorithm, transforming it into an interpretation. We considered the main activities performed by a student

together with parameters describing the activities. An example of an interpretation for a student is the following:

$$\left. \begin{aligned} & \{registration(par_1, \dots, par_n, 1), \\ & exam(par_1, \dots, par_m, 2), \\ & exam(par_1, \dots, par_m, 3), \\ & \dots \} \end{aligned} \right\}$$

where  $par_i$  means the  $i$ -th parameter for a certain activity.

A ten-fold cross-validation was used, i.e., the dataset was divided into ten sets (containing roughly the same proportion of positive and negative traces as the whole dataset) and ten experiments were performed, where nine sets were used for training and the remaining one for testing, i.e., for evaluating the accuracy of the learned theory. In particular, test sets contain either 33 positive and 49 negative traces or 32 positive and 48 negative traces. The same language bias was used in all ten experiments. The accuracy is defined as the number of compliant traces that are correctly classified as compliant by the learned model plus the number of non compliant traces that are correctly classified as not compliant divided by the total number of traces.

An IC theory is learned for each fold, composed of a number of rules between 25 and 31. The accuracy of the theories on the test sets ranges from 54% to 86%, with an average of 67.5%.

The second step was the assignment of weights to the ICs, by creating ten MLNs containing the theories translated into ML. Each of the ten MLNs were given as input to Alchemy for discriminative weight learning, which takes about 1.2 sec for every training set.

Ten MLN were also generated from the learned IC theories by assigning the pseudo-infinite weight  $10^{10}$  to all the clauses, in order to approximate a purely logical theory.

In the third step, we computed the probability of each test trace of being negative. This was performed by running the belief propagation inference algorithm of [21] (implemented in Alchemy) both on the MLNs with learned weights and on the MLNs with pseudo-infinite weights. In practice, we computed the marginal probabilities of the atoms of the form  $neg(i)$ , with  $i$  representing the identifier of a student in the test dataset.

Finally, we compared the sharp MLN with the weighted MLN using the average area under the ROC curve (AUC) [17] that has been identified as a better measure for evaluating the classification performances of algorithms with respect to accuracy, because it also takes into account the different distribution of positive and negative examples in the datasets. The sharp MLN achieved an average AUC of 0.7107528, while the weighted MLN achieved an average AUC of 0.7227286. We also applied a one-tailed paired  $t$  test: the null hypothesis that the two algorithms are equivalent can be rejected with a probability of 90.58%.

## 4 Related Works

Most works on process mining deal with process models in the form of graphs or Petri nets, that represent the allowed sequences of events as paths in the diagram. [5] presented an approach for inducing a process representation in the form of a directed graph encoding the precedence relationships. [4] proposed the  $\alpha$ -algorithm that induces Petri nets. The approach discovers binary relations in the log, such as the “follows” relation. The  $\alpha$ -algorithm is guaranteed to work for a restricted class of models. In [14] the result of induction is a process model in the form of a disjunction of special graphs called *workflow schemes*.

Recently, knowledge-based languages for the representation of process models have appeared. In them, models are seen as sets of constraints over the executions of the process. These models are more declarative because they state the conditions that process executions must satisfy rather than encoding them as paths in graphs.

Examples of declarative languages for representing process models are DecSerFlow [3], ConDec [2] and SCIFF [7,6]. [9] describes the relationships between these languages and shows that ConDec/DecSerFlow can be translated into SCIFF and a subset of SCIFF can be translated into ConDec/DecSerFlow.

[16] proposed the DPML algorithm that learns process models expressed in a subset of SCIFF. [15,8] presented the DecMiner system that is able to infer ConDec/DecSerFlow models by first inducing a SCIFF theory and then translating it into ConDec/DecSerFlow.

This paper extends the works [16,15,8] by including a probabilistic component in the process models. This allows to better model domains where the relationships among events are uncertain.

Recently, [20] discussed mining of process models in the form of AND/OR workflow graphs that are able to represent probabilistic information: each event is considered as a binary random variable that indicates whether the event happened or not and techniques from the field of Bayesian networks are used to build probability distribution over events. The paper presents a learning algorithm that induces a model by identifying the probabilistic relationships among the events from data. Thus the approach of [20] provides a probabilistic extension to traditional graph-based models, while we extend declarative modeling languages by relying on a first-order probabilistic language.

## 5 Conclusions

We propose a methodology, based on Statistical Relational Learning, for analyzing a log containing several traces of a process, labeled as compliant or non-compliant. From them we learn a set of declarative constraints expressed as ICs. Then we represent ICs in Markov Logic, a language extending first-order logic, to obtain a probabilistic classification of traces, by using the Alchemy system. Finally we evaluate the performances of the two models concluding that probabilistic ICs are more accurate than the pure logical ones. The experiments

have been performed on process traces belonging to a real dataset of university students' careers.

Supplementary material, including the code of the systems and an example dataset, can be found at <http://sites.google.com/a/unife.it/ml/pdpm/>.

## Acknowledgements

This work was possible thanks to the Audit Office of the University of Ferrara, in particular Alberto Domenicali and Susanna Nanetti, that supplied the university dataset for experiments.

## References

1. van der Aalst, W.M.P., van Dongen, B.F., Herbst, J., Maruster, L., Schimm, G., Weijters, A.J.M.M.: Workflow mining: A survey of issues and approaches. *Data Knowledge Engineering* 47(2), 237–267 (2003)
2. van der Aalst, W.M.P., Pesic, M.: A declarative approach for flexible business processes management. In: Eder, J., Dustdar, S. (eds.) *BPM Workshops 2006*. LNCS, vol. 4103, pp. 169–180. Springer, Heidelberg (2006)
3. van der Aalst, W.M.P., Pesic, M.: DecSerFlow: Towards a truly declarative service flow language. In: Bravetti, M., Núñez, M., Zavattaro, G. (eds.) *WS-FM 2006*. LNCS, vol. 4184, pp. 1–23. Springer, Heidelberg (2006)
4. van der Aalst, W.M.P., Weijters, T., Maruster, L.: Workflow mining: Discovering process models from event logs. *IEEE Transactions on Knowledge and Data Engineering* 16(9), 1128–1142 (2004)
5. Agrawal, R., Gunopulos, D., Leymann, F.: Mining process models from workflow logs. In: Schek, H.-J., Saltor, F., Ramos, I., Alonso, G. (eds.) *EDBT 1998*. LNCS, vol. 1377, pp. 469–483. Springer, Heidelberg (1998)
6. Alberti, M., Chesani, F., Gavanelli, M., Lamma, E., Mello, P., Torroni, P.: Verifiable agent interaction in abductive logic programming: The SCIFF framework. *ACM Transactions on Computational Logic* 9(4) (2008)
7. Alberti, M., Gavanelli, M., Lamma, E., Mello, P., Torroni, P.: An abductive interpretation for open societies. In: Cappelli, A., Turini, F. (eds.) *AI\*IA 2003*. LNCS, vol. 2829, Springer, Heidelberg (2003)
8. Chesani, F., Lamma, E., Mello, P., Montali, M., Riguzzi, F., Storari, S.: Exploiting inductive logic programming techniques for declarative process mining. In: Jensen, K., van der Aalst, W.M.P. (eds.) *Transactions on Petri Nets*. LNCS, vol. 5460, pp. 278–295. Springer, Heidelberg (2009)
9. Chesani, F., Mello, P., Montali, M., Storari, S.: Towards a decserflow declarative semantics based on computational logic. Technical Report DEIS-LIA-07-002, DEIS, Bologna, Italy (2007)
10. Clark, K.L.: Negation as failure. In: *Logic and Databases*. Plenum Press, New York (1978)
11. De Raedt, L., Van Laer, W.: Inductive constraint logic. In: Zeugmann, T., Shinohara, T., Jantke, K.P. (eds.) *ALT 1995*. LNCS (LNAI), vol. 997, Springer, Heidelberg (1995)



12. Domingos, P., Kok, S., Lowd, D., Poon, H., Richardson, M., Singla, P.: Markov logic. In: De Raedt, L., Frasconi, P., Kersting, K., Muggleton, S.H. (eds.) *Probabilistic Inductive Logic Programming*. LNCS (LNAI), vol. 4911, pp. 92–117. Springer, Heidelberg (2008)
13. Georgakopoulos, D., Hornick, M.F., Sheth, A.P.: An overview of workflow management: From process modeling to workflow automation infrastructure. *Distributed and Parallel Databases* 3(2), 119–153 (1995)
14. Greco, G., Guzzo, A., Pontieri, L., Saccà, D.: Discovering expressive process models by clustering log traces. *IEEE Transactions on Knowledge and Data Engineering* 18(8), 1010–1027 (2006)
15. Lamma, E., Mello, P., Montali, M., Riguzzi, F., Storari, S.: Inducing declarative logic-based models from labeled traces. In: Alonso, G., Dadam, P., Rosemann, M. (eds.) *BPM 2007*. LNCS, vol. 4714, pp. 344–359. Springer, Heidelberg (2007)
16. Lamma, E., Mello, P., Riguzzi, F., Storari, S.: Applying inductive logic programming to process mining. In: Blockeel, H., Ramon, J., Shavlik, J., Tadepalli, P. (eds.) *ILP 2007*. LNCS (LNAI), vol. 4894, pp. 132–146. Springer, Heidelberg (2008)
17. Provost, F.J., Fawcett, T.: Robust classification for imprecise environments. *Machine Learning* 42(3), 203–231 (2001)
18. Raedt, L.D., Dehaspe, L.: Clausal discovery. *Machine Learning* 26(2-3), 99–146 (1997)
19. Richardson, M., Domingos, P.: Markov logic networks. *Machine Learning* 62(1-2), 107–136 (2006)
20. Silva, R., Zhang, J., Shanahan, J.G.: Probabilistic workflow mining. In: Grossman, R., Bayardo, R.J., Bennett, K.P. (eds.) *Proceedings of the Eleventh ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 275–284. ACM, New York (2005)
21. Singla, P., Domingos, P.: Lifted first-order belief propagation. In: *Proceedings of the Twenty-Third AAAI Conference on Artificial Intelligence*, AAAI 2008, pp. 1094–1099. AAAI Press, Menlo Park (2008)

# Making Ontology-Based Knowledge and Decision Trees Interact: An Approach to Enrich Knowledge and Increase Expert Confidence in Data-Driven Models

Iyan Johnson<sup>1,3</sup>, Joël Abécassis<sup>1</sup>, Brigitte Charnomordic<sup>3</sup>,  
Sébastien Destercke<sup>1,\*</sup>, and Rallou Thomopoulos<sup>1,2</sup>

<sup>1</sup> IATE Joint Research Unit, UMR1208, CIRAD-INRA-Supagro-Univ. Montpellier II  
2 place P. Viala, F-34060 Montpellier cedex 1

<sup>2</sup> LIRMM, CNRS-Univ. Montpellier II, 161 rue Ada, F-34392 Montpellier cedex 5

<sup>3</sup> INRA, UMR 729 MISTEA, F-34060 Montpellier, France  
destercke@supagro.inra.fr

**Abstract.** When using data-driven models to make simulations and predictions in experimental sciences, it is essential for the domain expert to be confident about the predicted values. Increasing this confidence can be done by using interpretable models, so that the expert can follow the model reasoning pattern, and by integrating expert knowledge to the model itself. New pieces of useful formalised knowledge can then be integrated to an existing corpus while data-driven models are tuned according to the expert advice. In this paper, we propose a generic interactive procedure, relying on an ontology to model qualitative knowledge and on decision trees as a data-driven rule learning method.

A case study based on data issued from multiple scientific papers in the field of cereal transformation illustrates the approach.

## 1 Introduction

In many domains where extensive mathematical knowledge is not available, sharing expertise and conclusions obtained from data are of great importance for building efficient decision support tools. This is very much the case in Life Sciences [1], owing to the great variability of living organisms and to the difficulty of finding universal deterministic natural laws in biology. Many areas of life science (food processing, cultural practices, transformation processes) rely as much upon expertise and data than upon mathematical models.

For domain experts to use data-driven models (especially in sciences where experiments play a central role), it is necessary for them to be confident in the results. Even if confidence can be partially obtained by a numerical validation procedure, an expert will rely more on the results if he/she can understand the prediction basis and if the reasoning uses knowledge of the (natural or industrial) processes and of their interactions. This can be done by using interpretable rule learning models, such as decision trees, fuzzy rule bases, Bayesian networks, . . . Unfortunately, experimental data are seldom collected with a global approach, i.e. with the thought that they are only a part

---

\* Corresponding author.

of a more complex system, and are not usually ideally structured to achieve inductive learning. Learning models from rough experimental data therefore seldom provides domain experts with completely meaningful and sensible models. A review of results of interpretable data-driven models whose descriptive variables do not necessarily exactly coincide with the ones domain experts would have initially selected has a double benefit. First it can be a means to acquire new items of knowledge from experts, then it is a good way to design a useful model.

In this paper, we propose an interactive (between AI methods and domain experts) and iterative approach to achieve these two related goals which are usually hard to fulfill, i.e. enrich our qualitative knowledge of processes and increase the expert confidence in the data-driven model.

Domain knowledge (expert interviews, literature, ...) is formalized by using an ontology to specify a set of concepts and the relations linking them, which gives a structure that facilitates the interaction with domain experts. Our approach is generic regarding data-driven rule learning methods, and in the following, we illustrate it with decision trees. Decision tree algorithms are efficient approaches for data-driven discovery of complex and non obvious relationships. Their readability and the absence of *a priori* assumptions explain their popularity. They are particularly useful for variable selection in highly multidimensional problems, therefore they are ideal to display statistically important variables on which the domain expert should focus. Decision trees can be pruned and, as thoroughly discussed in [2], not too complex. Such a low complexity is essential for the model to be interpretable, as confirmed by the conclusions of Miller ([3]) relative to the *magical number* seven.

As far as we know, no interactive approach trying to combine qualitative knowledge (modelled by an ontology) and data-driven rule learning methods in the field of experimental sciences has been proposed up to now. Indeed, most attempts at such collaborative methods focus on problems where scalability is a main issue, and where method performances can be automatically measured. A few semi-automatic interactive approaches (combining learning and ontology-based knowledge) recently appeared in the literature, in fields where large amounts of data must be treated, such as the Semantic Web ([4],[5]), to deal with multiple ontologies ([6],[7]), or in cases where data are well-structured, such as in image classification ([8]).

The case of inductive learning using ontologies, data and decision trees has been addressed in [9], however it is limited to the specific case of taxonomies<sup>1</sup>, whereas in this paper we do not make this restriction. Moreover we consider domain expert knowledge and feedback, while the approach in [9] is better for fully automatic learning (once the ontology is given).

In many cases in Life Sciences, data can be scarce, costly, and not necessarily numerous. Our purpose is to propose a framework to make the best of these data. Therefore our primary aim is to not to improve the numerical accuracy of a learnt model (although it is certainly desired), or the fastness with which it detects some features.

We are aware of the challenge to achieve a good balance between the time spent by the domain expert on the learning task and the benefits he can retrieve in terms of model generalization and fiability. Our purpose is to tend towards automated procedures

---

<sup>1</sup> Ontologies that can be represented as rooted trees in graph theory.

as much as possible, where domain experts, ontology and learning models can interact without the help of AI experts.

The paper is organized as follows: Section 2 provides the background and definitions necessary to understand the paper. Section 3 formally describes the various data processing operations done using the ontology. Section 4 presents the outline of the interactive approach. A case study concerning the impact of agri-food transformation processes on the nutritional quality of wheat-based products is presented in section 5. All along the paper, we illustrate our generic approach by taking examples in the field of expert knowledge, scientific papers and experiments related to cereal product quality.

## 2 Background

In this section, we briefly recall essential elements regarding ontology definition and decision trees, which will be used as data-driven inductive learning methods to provide domain expert readable models.

### 2.1 Ontology Definition

The ontology  $\Omega$  is defined as a tuple  $\Omega = \{\mathcal{C}, \mathcal{R}\}$  where  $\mathcal{C}$  is a set of concepts and  $\mathcal{R}$  is a set of relations.

**Relationship between concepts and variables.** We consider a data set  $\mathbb{D}$  containing  $K$  variables and  $N$  experiments. Each variable  $X_k$ ,  $k = 1, \dots, K$ , is a concept  $c \in \mathcal{C}$  in the ontology  $\Omega$ . The  $n^{th}$  value of the  $k^{th}$  variable is denoted  $x_{k,n}$ .

**Concept range.** A concept  $c$  may be associated with a definition domain by the *Range* function. This definition domain can be: (i) *numeric*, i.e.  $Range(c)$  is a closed interval  $[min_c, max_c]$ ; (ii) *'flat' (non hierarchized) symbolic*, i.e.  $Range(c)$  is an unordered set of constants, such as a set of scientific papers; (iii) *hierarchized symbolic*, i.e.  $Range(c)$  is a set of partially ordered constants, themselves concepts belonging to  $\mathcal{C}$ .

**Set of relations.** The set of relations  $\mathcal{R}$  is composed of:

- a) the *subsumption* or 'kind of' relation denoted by  $\preceq$ , which defines a partial order over  $\mathcal{C}$ . Given  $c \in \mathcal{C}$ , we denote by  $\mathcal{C}_c$  the set of sub-concepts of  $c$ , such that:  $\mathcal{C}_c = \{c' \in \mathcal{C} | c' \preceq c\}$ . When  $c$  represents a variable with hierarchized symbolic definition domain, we have  $Range(c) = \mathcal{C}_c$ . For brevity, we shall use  $\mathcal{C}_c$  in the sequel whenever possible.
- b) a set of *functional dependencies*. A functional dependency  $FD$  expresses a constraint between two sets of variables and is represented as a relation between two sets of concepts of  $\mathcal{C}$ . Let  $X = \{X_{k_1}, \dots, X_{k_2}\} \subseteq \mathcal{C}$ ,  $1 \leq k_1 \leq k_2 \leq K$  and  $Y = \{Y_{k_3}, \dots, Y_{k_4}\} \subseteq \mathcal{C}$ ,  $1 \leq k_3 \leq k_4 \leq K$  be two disjoint subsets of concepts.  $X$  is said to functionally determine  $Y$  if and only if there is a function  $DetVal_{FD}$  such that:  $DetVal_{FD} : Range(X_{k_1}) \times \dots \times Range(X_{k_2}) \rightarrow Range(Y_{k_3}) \times \dots \times Range(Y_{k_4})$ .

Two instances of such functional dependencies are required in our approach:

1. a *property* relation  $\mathcal{P} : \mathcal{C} \rightarrow 2^{|\mathcal{C}|}$  that maps a single concept to a set of other concepts that represent associated properties.

*Example 1.*  $\mathcal{P}(\text{Vitamin}) = \{\text{Thermosensitivity, Solubility, } \dots\}$ .

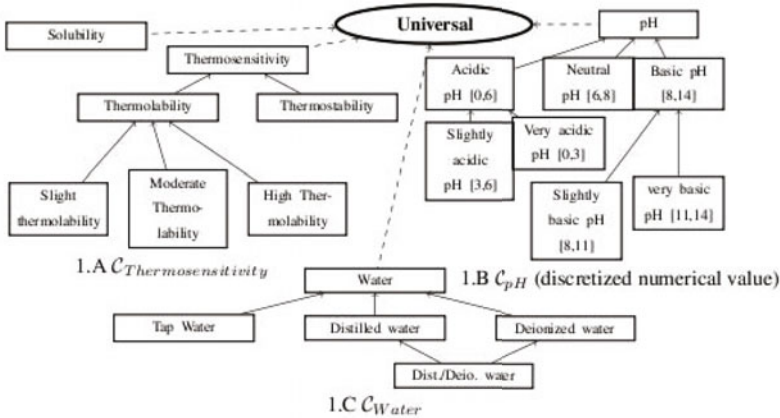
For each concept that has some properties, i.e.,  $\forall c \in \mathcal{C}, \mathcal{P}(c) \neq \emptyset$ , we denote by  $p_c$  the number of properties and by  $\mathcal{P}(c)_i$  the  $i$ th element of  $\mathcal{P}(c)$ , with  $i = 1, \dots, p_c$ . The function  $DetVal_{\mathcal{P}}$  will be denoted by  $\mathcal{HP}_c$  (for *HasProperty*). It maps a particular value of  $Range(c)$  to the particular property values it takes in the ranges of the concepts of  $\mathcal{P}(c)$ .  $\mathcal{HP}_c : Range(c) \rightarrow Range(\mathcal{P}(c)_1) \times \dots \times Range(\mathcal{P}(c)_{p_c})$ . We denote by  $\mathcal{HP}_{c \downarrow i} : Range(c) \rightarrow Range(\mathcal{P}(c)_i)$  the restriction of  $\mathcal{HP}_c$  to its  $i$ th property, that is  $\mathcal{HP}_{c \downarrow i} = \mathcal{HP}_c \cap (Range(c) \times Range(\mathcal{P}(c)_i))$ .

*Example 2.* We have  $\mathcal{P}(\text{Vitamin})_1 = \text{Thermosensitivity}$ .

2. a *determines* relation  $\mathcal{D} : 2^{|\mathcal{C}|} \rightarrow \mathcal{C}$  which specifies a subset of concepts whose values entirely determine the value taken by another concept.

*Example 3.*  $\mathcal{D}(\{\text{Pastatype, Cookingtime}\}) = \text{Cookingtype}$  models the fact that the *Cooking type* is a function depending on the values of *Pasta type* and of *Cooking time*.

The function  $DetVal_{\mathcal{D}}$  will be denoted by  $\mathcal{HD}_C$  (for *HasDetermination*).  $\forall C \in 2^{|\mathcal{C}|}$  such that  $\mathcal{D}(C) \neq \emptyset$ , we define the function  $\mathcal{HD}_C$  such that  $\mathcal{HD}_C : Range(c_1) \times \dots \times Range(c_{|C|}) \rightarrow Range(\mathcal{D}(C))$ , with  $c_i$  and  $|C|$  being respectively the  $i$ th element and the number of elements of  $C$ . The function  $\mathcal{HD}$  simply gives the values of  $\mathcal{D}(C)$ , given the values of the determinant variables.



**Fig. 1.** Some variables and related ontology parts where  $A \rightarrow B$  means that  $A$  is a kind of  $B$

*Example 4.*  $\mathcal{HD}(\{Short, 18min\}) = Overcooking$ .

Figure 1 gives an example of three categorical variables: *pH*, *Water* and *Thermosensitivity*, together with the sub-ontologies induced by the order  $\preceq$ . *pH* is an example of a continuous variable discretized into a categorical variable. Note that  $\mathcal{C}_{Water}$  is *not* a simple taxonomy. We will repeatedly refer to this figure in our forthcoming examples.

## 2.2 Decision Trees

Decision trees are well established learning methods in supervised data mining. They can handle both classification and regression tasks. In multidimensional modeling, they perform well in attribute selection and are often used prior to further statistical modeling. Also note that decision trees algorithms include methods to deal with missing data, meaning that every experiment (or data), even the one with lacking values for some variables, is used in the process. In this paper, due to lack of space, we focus on the C4.5 [10] family of decision trees, and we use them for classification. In the present study, another main interest of decision trees are their interpretability by domain experts, due to their graphical nature.

**Algorithm description.** Input to classification decision trees consists of a collection of training cases, each having a tuple of values for a set of input variables, and a discrete output variable  $Y$  divided into  $M_Y$  classes:  $(\mathbf{x}_n, \mathbf{y}_n) = (x_{1,n}, x_{2,n} \dots x_{K,n}, y_n)$ . An attribute  $X_k$  can be continuous or categorical. The goal is to learn from the training cases a recursive structure (taking the shape of a rooted tree) consisting of (i) leaf nodes labeled with a class value, and (ii) test nodes (each one associated to a given variable) that can have two or more outcomes, each of these linked to a subtree.

Well-known drawbacks of decision trees are the sensitivity to outliers and the risk of over-fitting. To avoid over-fitting, cross-validation is included in the procedure and to gain in robustness, a pruning step usually follows the tree growing step (see [11,10]).

**Splitting criteria.** We denote by  $p_m(S)$  the proportion of examples at node  $S$  that belong to class  $m$ . To select the splitting variable at node  $S$ , the C4.5 algorithm examines all candidate variables in turn, and computes the potential improvement brought by each of them. It then selects the variable that yields the best improvement.

Let us denote by  $M_k$  the number of modalities of  $X_k$ . The improvement gained by splitting the node  $S$  into  $M_k$  subsets  $S_1, S_2 \dots S_{M_k}$  according to  $X_k$ , is evaluated as

$$G(S, X_k) = I(S) - \sum_{i=1}^{M_k} \frac{|S_i|}{|S|} I(S_i)$$

with  $M_k$  the number of possible outcomes.

$I(S)$  is derived from the information theory entropy, its value at node  $S$  is

$$I(S) = - \sum_{m=1}^{M_Y} p_m(S) \log_2 p_m(S).$$

### 3 Data Processing Using Ontologies

When automatically processing data to perform knowledge discovery or classification, some input variables and/or their modalities may be irrelevant to the problem under study. Indeed, experimental data reported in papers, reports, etc., are usually collected for specific research objectives and may not entirely fit in a global knowledge engineering approach. In some cases, a particular variable may be decomposed into some properties more significant for the expert. For instance, to appreciate the degradation of vitamin component during the *Cooking in water* operation, it is better to consider the vitamin thermosensitivities and carbonate reactivities rather than the vitamin types. Also, the variable modalities may be too numerous, and a noise source. For example, a pH value may be divided into *slightly*, *moderately*, *very acid* and *basic*, whereas separating between *acid* and *basic* pH is sufficient.

This section details various data transformations exploiting both the ontology defined in Sect. 2.1 and domain expert feedbacks to build more significant variables from the original ones. These transformations are performed automatically, according to the used ontological knowledge (note that this ontological knowledge may not be available initially). Transformed data can then be re-used in the learning process, thus providing a new model. Feedbacks may be stimulated by a third-party data treatment method, i.e., decision trees in the present paper. Appropriate transformations are selected by an expert evaluation of learning results.

#### 3.1 Replacement of a Variable by New Ones

This process consists of substituting a variable by some of its (more relevant) properties, which then become new variables. Let  $X_k$  be a variable such that  $\forall n \in [1; N], \mathcal{P}(X_k) \neq \emptyset$ . For each property  $\mathcal{P}(X_k)_i, i \in [1; p_{X_k}]$  (or a subset of them), we create a new variable  $X_{K+i}$  such that:  $\forall n \in [1; N] \quad x_{K+i,n} = \mathcal{HP}_{X_k}(x_{k,n})_{\downarrow i}$ , with  $\mathcal{HP}_{X_k}(x_{k,n})_{\downarrow i}$  the projection of  $\mathcal{HP}_{X_k}(x_{k,n})$  on  $\text{Range}(\mathcal{P}(X_k)_i)$  and  $\mathcal{P}(X_k)_i$  the  $i$ th element of  $\mathcal{P}(X_k)$ . Indeed, a given variable may summarize many aspects of a process, and it is sometimes desirable to decompose it into influential properties (for example, the "year effect" often considered in crop management summarizes information related to temperatures, climatic conditions, presence of diseases, ...).

*Example 5.* Let  $X_k = \textit{vitamin}$  be the (non relevant) variable to be replaced and  $\mathcal{P}(\textit{vitamin}) = \{\textit{solubility}, \textit{thermosensitivity}\}$  its properties. We have  $X_{K+1} = \textit{solubility}$  and  $X_{K+2} = \textit{thermosensitivity}$ . The new variables are *solubility* and *thermosensitivity*. Now, if for the  $n$ th experiment,  $x_{k,n} = \textit{VitaminA}$ , the two new values for the  $n$ th experiment are  $x_{K+1,n} = \mathcal{HP}_{X_k}(x_{k,n})_1 = \textit{Liposoluble}$  and  $x_{K+2,n} = \mathcal{HP}_{X_k}(x_{k,n})_2 = \textit{Thermolability}$ . The initial variable  $X_k = \textit{Vitamin}$  is removed.

#### 3.2 Grouping the Modalities of a Variable Using Common Properties

In some cases, it may be useful to consider subsets of modalities corresponding to a particular feature rather than the modalities themselves. Formally, this is equivalent to considering elements of the power set of modalities, these elements being chosen w.r.t.

some properties of the variable. Let  $X_k$  be a given variable such that  $\mathcal{P}(X_k) \neq \emptyset$  and let  $i \in [1; p_{X_k}]$ . We replace  $X_k$  by  $X'_k$  such that, for  $n \in [1; N]$ :

$$z_n = \mathcal{HP}_{X_k}(x_{k,n})_{\downarrow i}, z_n \in \text{Range}(\mathcal{P}(X_k)_i) \quad \text{and} \quad x'_{k,n} = \mathcal{HP}_{X_k}^{-1}(z_n).$$

The first equation expresses that we first get  $z_n$ , the  $i$ th property value associated with  $x_{k,n}$ . The second equation expresses the search for all the antecedents, i.e. all  $x_{k,l}$  ( $l \in [1; N]$ ) whose  $i$ th property value is equal to  $z_n$ , which includes  $x_{k,n}$  but may also include other values.

*Example 6.* Let  $X_k = \text{Water}$  and  $pH \in \mathcal{P}(\text{Water})$ . Suppose that we want to keep track of the types of water used in the experiments, but that it would be desirable to group them by  $pH$ . We have  $\mathcal{HP}_{\text{Water}}(\text{Tap water})_{\downarrow pH} = \text{Basic pH}$ , and  $\mathcal{HP}_{\text{Water}}(c)_{\downarrow pH} = \text{Neutral pH}$  for any other  $c \in \mathcal{C}_{\text{Water}}$ . The new variable  $X'_k$  thus has the following two modalities:  $\{\text{Tap Water}\}$  and  $\{\text{Deionized water, Distilled water, Distilled deionized water}\}$ . Since the second modality is multi-valued, it can then be replaced by a new concept *Ion-poor water* in  $\mathcal{C}$ , added as a sub-concept of *Water* and a super-concept of *Distilled water* and *Deionized water* (see Fig. 1).

### 3.3 Merging of Variables in Order to Create a New One

It may be relevant to merge several variables into another variable, with the values of the latter defined by the values of the former. It both facilitates the interpretation (as less variables are considered) and avoids to consider as significant a single variable that is only significant (at least from an expert standpoint) in conjunction with other variables. Let  $C = \{X_{k_1}, \dots, X_{k_{|C|}}\} \in 2^{\mathcal{X}}$  such that  $\mathcal{D}(C) \neq \emptyset$ .

Then we define a new variable:  $\mathcal{X}_{K+1} = \mathcal{D}(\{X_{k_1}, \dots, X_{k_{|C|}}\})$  such that:

$$\forall n \in [1; N] \quad x_{K+1,n} = \mathcal{HD}_C(\{x_{k_1,n}, \dots, x_{k_{|C|},n}\}).$$

*Example 7.* When cooking pasta, domain experts differentiate between *Under-cooked*, *Over-cooked*, and *Optimally cooked* products. However, these states depend on the type of pasta and on the cooking-time, which are usually measured in experiments. Therefore, it makes sense to replace *Cooking time* and *Pasta type* by a new variable *Cooking type*. For example,  $\mathcal{HD}_C(\{18\text{min}, \text{Short}\}) = \text{Over-cooked}$ , replaces *Cooking time*=18 and *Pasta type*=*Short* with *Over-cooked*, in all relevant experiments.

## 4 Interactive Approach: Principles and Evaluation

In this section, we first present the principles of our interactive approach. Then we detail the way we shall evaluate the approach and its results.

### 4.1 Principles

We assume that we start from an initial domain ontology  $\Omega_0 = \{\mathcal{C}_0, \mathcal{R}_0\}$ , that can be obtained from semi-automated methods [12], domain expert elicitation or that is readily available. We also assume that an initial learning data set  $\mathbb{D}_0$  is available, whose variables coincide with the ontology concepts.



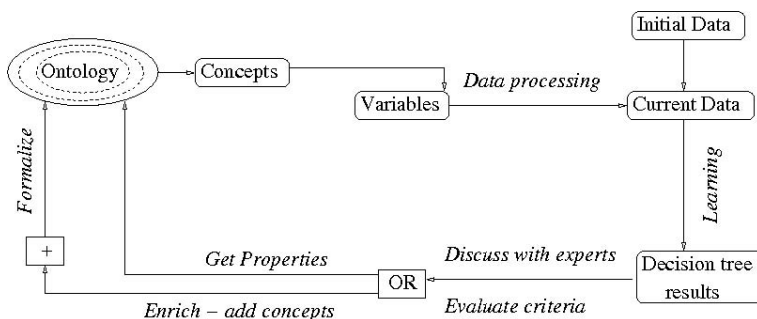


Fig. 2. Collaborative method scheme

Learning methods and ontology-based knowledge are combined through an interactive and iterative process represented in Figure 2. At step  $i$ , it can be summarized as follows:

1. Induce model  $\mathbb{M}_i$  from data, using the data set  $\mathbb{D}_i$  (starting with  $\mathbb{D}_0$ );
2. Assess numerical accuracy of  $\mathbb{M}_i$  and discuss its significance with domain experts;
3. If domain experts are satisfied, stop the process, if not, elicit from domain expert the transformations to be done on variables, as well as the modalities, properties or functional dependencies used in this transformation. Add newly identified concepts and relations to the ontology  $\Omega_i$ , obtaining  $\Omega_{i+1}$  (starting with  $\Omega_0$ );
4. Using  $\Omega_{i+1}$  and domain expert opinion, transform data (using methods from Sect. 3) to obtain  $\mathbb{D}_{i+1}$  from  $\mathbb{D}_i$ ;
5. Set  $i = i + 1$  and go back to step 1;

## 4.2 Evaluation

There are two ways in which the current method can be evaluated:

- *subjective* human evaluation, performed by domain experts assessing their confidence in the results, and which potential inconsistencies they detected in the model,
- *objective* automatic numerical evaluation, where the results and stability of the predictive models are measured by numerical indices.
  - The classical criterion for classification trees is the misclassification rate,  $Ec = \frac{MC}{N}$ , where MC is the number of misclassified items and  $N$  is the data set size, computed with a cross validation procedure or on the whole data set.
  - Tree complexity:  $Nrules + Nnodes/Nrules$ , where  $Nrules$  is the number of terminal nodes (leaves), which is equivalent to the number of rules, and  $Nnodes$  is the total number of nodes in the tree.

During evaluation, it is also important to take into consideration the data fiability: sources, experimental equipment, protocol . . .

## 5 Case Study: Application to Food Quality Prediction

Cereal and pasta industry has developed from traditional companies relying on experience and having a low rate of innovation, to a dynamic industry geared to follow consumer trends: healthy, safe, easy to prepare, pleasant to eat [13].

Previous systems have been proposed in food science, and more specifically in the field of cereal transformation, in order to help prediction [14]. However none of them takes into account both experimental data and expert knowledge, nor proposes solutions in absence of a predetermined (mathematical or expert) model.

### 5.1 Context and Description of the Case Study

For each unit operation of the transformation process, and for each family of product properties, information is given as a data set. The input variables are the operation parameters. The output variable is the operation impact on a property (e.g. the variation of vitamin content). Here, we study the case of the *Cooking in water* unit operation and the *Vitamin content* property. This case concerns 150 experimental data and involves 60 of the ontology concepts. Table 1(a) shows some values of the input variables and of the output variable. The ontology was created using CoGUI (<http://www.lirmm.fr/cogui/>). Data transformation and decision trees were obtained using the R software [15] (use of *R-WEKA* package and about 2000 lines of developed code).

**Table 1.** (a) Part of the training data set (b) Tree evaluation

Id	Vitamin	Cooking temp. (C)	Cooking time (min)	Water	Vitamin loss (%)
1	B6	100	13	NA	-52
2	B2	100	12	Tap	-53
3	B1	98	15	Distilled	-47
4	B2	90	10	NA	-18
5	B1	100	NA	Dist./Deio.	-41

Iteration #	MC rate (%)	Complexity
1	44	7.3
2	48	8.4
3	35	7.5
4	35	7.5

### 5.2 Application of the Approach to the Case Study

The approach has been carried out with a strong collaboration between a team of four computer science researchers and two food science researchers<sup>2</sup>, with a regular involvement of all participants. The output variable is the *Percentage of vitamin loss* during the process, which is a continuous variable, discretized into four ordered classes *Low loss*, *Average loss*, *High loss*, *Very high loss*.

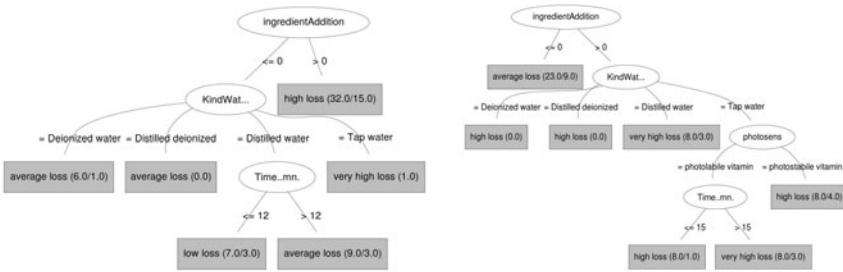
The implementation used for decision trees is the R software with the *R-WEKA* package. All trees are built using a minimum number of instances per leaf equal to 6, and then pruned. The plots are to be interpreted as follows:

<sup>2</sup> B. Cuq (Prof. in Food Science), J. Abécassis (Research Eng. in Cereal Technology), IATE Joint Research Unit.

1. Each test node is labeled by the splitting variable.
2. for each leaf node, the number of misclassified observations is specified.

Our approach will be conforming to the iterative approach outlined in section 4. It will be illustrated by four iterations.

**Iteration 1: initial state.** Figure 3 shows the tree trained on the raw data sample ( $\mathbb{D}_0$ ). As mentioned in Section 2.2, the complexity for C4.5 decision trees increases with the number of modalities, which is the case for the *Kind of water* variable. The purpose of our approach is also to reduce that complexity by identifying the relevant underlying properties hidden behind these modalities.



**Fig. 3.** Decision trees on - (a) raw data - (b) data with vitamin properties

Examination of the tree by domain experts led to the following remarks and adjustments. First, the most discriminant variable is *Ingredient Addition*. Indeed, it corresponds to adding vitamins for compensating a loss during the cooking process. The experts suggested to *enrich the ontology* by characterizing the vitamins by their properties. The following elements were added to the ontology (obtaining  $\Omega_1$ ), and data were transformed to obtain  $\mathbb{D}_1$ .

$$\mathcal{P}(Vitamin) = \{Solubility, Thermosensitivity, Photosensitivity, \dots\}$$

$$Range(Photosensitivity) = \{Photolabile, Photostabile\}$$

$$\mathcal{HP}_{Vitamin}(VitaminA) = \{Liposoluble, Thermolabile, Photostabile\}$$

**Iteration 2: introducing knowledge on Vitamin properties.** The model  $\mathbb{M}_1$  is a new tree illustrated by Figure 3(b). The *Kind of water* and the *Cooking time* variables are emphasized by this tree. And yet, discussion with experts brought out the fact that the *Cooking time* variable is relevant only if considered with the *Pasta type*. Experts also suggested that water can be better characterized in terms of *pH* and of *Hardness*. In the available experiments the water *pH* and *Hardness* were not measured. However they can be reconstructed from the water types. The following elements were added to  $\Omega_1$  to obtain  $\Omega_2$  and used to transform  $\mathbb{D}_1$  in  $\mathbb{D}_2$  (see section 2.1):

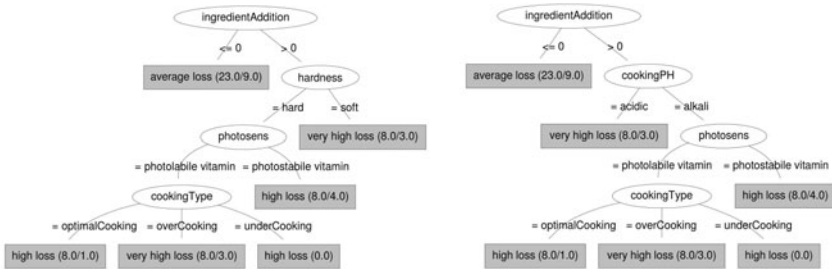
$$\begin{aligned} \mathcal{P}(\text{Water}) &= \{pH, \text{Hardness}\}, & \text{Range}(ph) &= \{\text{AcidpH}, \text{NeutralpH}, \text{BasicpH}\} \\ \mathcal{HP}_{\text{water}}(\text{Tapwater}) &= \{\text{NeutralpH}, \text{Hard}\} \\ \mathcal{D}(\{\text{Pastatype}, \text{Cookingtime}\}) &= \text{Cookingtype}, & \mathcal{HD}(\{\text{short}, 18\text{min}\}) &= \text{Overcooking} \end{aligned}$$

**Iteration 3: introducing Cooking type and Water properties.** Figure 4(a) shows  $\mathbb{M}_2$ , the tree obtained with the previous modifications. We can see on this tree that the *Hardness*, the newly built variable, is now selected for the second split. The discussion with experts highlighted the existence of a link between *Water hardness* and *pH* evolution. The water pH evolution depends both on the *Cooking temperature* and on the *Water hardness*. A new variable will then be created according to a few expert rules not detailed here, obtaining  $\Omega_3$  and  $\mathbb{D}_3$ .

$$\mathcal{D}(\{pH, \text{Temperature}\}) = \text{CookingpH}$$

**Iteration 4: introducing the Cooking pH.** Figure 4(b) displays the final C4.5 tree (model  $\mathbb{M}_3$ ). Relevant variables are now selected by the learning algorithm. In particular, some initially measured continuous variables, such as *Cooking time*, are now replaced by more meaningful ones, such as *Cooking type*, which is obtained upon adding a new concept to the ontology, i.e. *Pasta type*.

Table 1(b) presents the evolution of the criteria defined in section 4.2. Though the misclassification rate remains high, essentially due to the data scarcity, it is better for the last two iterations, while the complexity remains low. Further investigation, through the examination of the confusion matrix, showed that almost all prediction errors are due to the assignment of a label *close* to the *right* one, for instance *High Loss* instead of *Very High Loss*.



**Fig. 4.** Decision tree - (a) including Cooking Type and Water Properties - (b) at the final step

## 6 Conclusion

Formalizing and acquiring new expert knowledge, as well as the construction of reliable models are two important aspects of artificial intelligence research in experimental

sciences. Of particular importance is the confidence that domain experts grant to statistically learnt models. As in other domains (e.g., the semantic web), both data-driven and ontological knowledge can help each other in their respective tasks.

In this paper, we proposed a collaborative and iterative approach, where expert knowledge and opinion issued from learnt models was integrated to the ontology describing the domain knowledge. This formalization is then re-used to transform available data and to learn new models from them, these new models being again the source of additional expert opinions, and so on until experts are satisfied with the results. This allows both to enrich the ontological knowledge and to increase expert confidence in the results delivered by learning methods.

The proposed approach is applied to a case study in the field of cereal transformation. This case study was undertaken iteratively, in tight collaboration with domain experts. It demonstrates the added value of taking into account ontology-based knowledge, by improving the result interpretability and relevance. It also aims to extract, by presenting expert with data-driven models, ontological knowledge that may be useful in other applications.

The present work is a first step to meet the difficult challenge of building semi-automated methods. There are several perspectives for future work in that direction: to handle missing (or imprecisely defined) items in a more appropriate way (for instance using imprecise probabilities as in recent approaches, see [16]); to consider instances whose properties are only partially known; to define new tree evaluation criteria regarding the stability of the selected variables; to automate the whole process so that AI expert are not needed to perform the analysis.

## References

1. Seising, R.: Soft computing and the life science-philosophical remarks. In: IEEE International Conference on Fuzzy Systems, July 2007, pp. 798–803. IEEE, Los Alamitos (2007)
2. Ben-David, A., Sterling, L.: Generating rules from examples of human multiattribute decision making should be simple. *Expert Syst. Appl.* 31(2), 390–396 (2006)
3. Miller, G.A.: The magical number seven, plus or minus two: Some limits on our capacity for processing information. *Psychological Review* 63, 81–97 (1956)
4. Stumme, G., Hotho, A., Berendt, B.: Semantic web mining: State of the art and future directions. *J. of Web Semantics* 4, 124–143 (2006)
5. Adomavicius, G., Tuzhilin, A.: Expert-driven validation of rule-based user models in personalization applications. *Data Mining and Knowledge Discovery* 5(1-2), 33–58 (2001)
6. Parekh, V., Gwo, J.P.J.: Mining Domain Specific Texts and Glossaries to Evaluate and Enrich Domain Ontologies. In: International Conference of Information and Knowledge Engineering, Las Vegas, NV, The International Multi Conference in Computer Science and Computer Engineering (June 2004)
7. Ling, T., Kang, B.H., Johns, D.P., Walls, J., Bindoff, I.: Expert-driven knowledge discovery. In: Latifi, S. (ed.) *Proceedings of the Fifth International Conference on Information Technology: New Generations*, pp. 174–178 (2008)
8. Maillot, N., Thonnat, M.: Ontology based complex object recognition. *Image and Vision Computing* 26, 102–113 (2008)
9. Zhang, J., Silvescu, A., Honavar, V.: Ontology-driven induction of decision trees at multiple levels of abstraction. *LNCS*, pp. 316–323 (2002)

10. Quinlan, J.: C4. 5: programs for machine learning. Morgan Kaufmann, San Francisco (1993)
11. Quinlan, J.: Induction of decision trees. *Machine learning* 1(1), 81–106 (1986)
12. Thomopoulos, R., Baget, J., Haemmerle, O.: Conceptual graphs as cooperative formalism to build and validate a domain expertise. In: Priss, U., Polovina, S., Hill, R. (eds.) ICCS 2007. LNCS (LNAD), vol. 4604, p. 112. Springer, Heidelberg (2007)
13. Dalbon, G., Grivon, D., Pagnani, M.: Continuous manufacturing process. In: Kruger, J., Matsuo, R., Dick, J. (eds.) *Pasta and Noodle Technology*, AACC, St Paul, MN-USA (1996)
14. Young, L.: Application of Baking Knowledge in Software Systems. In: *Technology of Bread-making*, 2nd edn., pp. 207–222. Springer, US (2007)
15. R Development Core Team: *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria (2009) ISBN 3-900051-07-0
16. Strobl, C.: *Statistical Issues in Machine Learning - Towards Reliable Split Selection and Variable Importance Measures*. PhD thesis, Ludwig-Maximilians-University Munich, Germany (2008)

# A Novel Initialization Method for Semi-supervised Clustering

Yanzhong Dang, Zhaoguo Xuan, Lili Rong, and Ming Liu

Institute of Systems Engineering, Dalian University of Technology, Dalian, 116024, China  
yzhdang@dlut.edu.cn, xzg@dl.cn, llrong@dlut.edu.cn,  
mingliu@mail.dlut.edu.cn

**Abstract.** In recent years, the research of semi-supervised clustering has been paid more and more attention. For most of the semi-supervised clustering algorithms, a good initialization method can create the high-quality seeds which are helpful to improve the clustering accuracy. In the real world, there are few labeled samples but many unlabeled ones, whereas most of the existing initialization methods put the unlabeled data away for clustering which may contain some potentially useful information for clustering tasks. In this paper, we propose a novel initialization method to transfer some of the unlabeled samples into labeled ones, in which the neighbors of labeled samples are identified at first and then the known labels are propagated to the unlabeled ones. Experimental results show that the proposed initialization method can improve the performance of the semi-supervised clustering.

**Keywords:** Initialization method, semi-supervised clustering, similarity measure.

## 1 Introduction

In the real-world applications of data mining, there are normally abundant unlabeled samples. On the other hand, the labeled samples are scarce, which are costly to obtain since making examples labeled would require much more human efforts. With a few labeled samples for clustering, it is known as the semi-supervised clustering problem which has recently been studied by many researchers with great interests [1].

As a part of semi-supervised clustering algorithm, the initialization of data items plays an important role in improving the quality of clusters, especially for the iterative refinement clustering. Such problem has been studied in different ways. In [2], Basu proposed an initialization method which used labeled samples to form the initial centroids as the seeds of clusters. But they did not consider the incomplete label situation which implied that the labeled samples could not cover all the categories in most cases. Zhong [3] considered the information on unlabeled samples but only selected the labeled samples as the seeds. When no labeled samples could be used, he applied unlabeled samples as the seeds by means of the modified KKZ (firstly proposed by Katsavounidis, Kuo and Zhang [4]). Sun et al. [5] presented an extreme way to the unlabeled samples. If samples in a certain cluster are all unlabeled, they can be ignored by decreasing the current number of clusters. The initialization

methods mentioned above did not make good use of the unlabeled data samples which might contain some meaningful information for clustering. However, the reality is the labeled data is often short, sometimes incomplete, and the unlabeled data is abundant. How to use the whole data samples no matter labeled or not to form the seeds of clusters is the key problem for the initialization process of semi-supervised clustering methods.

In the paper, we propose a novel initialization method which makes full use of data information by propagating the given labels to more unlabeled data samples. Based on the assumption presented by Zhu [6], i.e. the approximate data samples likely have the same label, we therefore calculate the similarity between the given labeled samples and their neighbors without labels. If they are similar enough, the samples without labels could become the candidates to be propagated in terms of a predetermined threshold. Once the candidates are assigned the labels, more labeled data samples can be reached which are helpful to find more correct seeds of clusters. Experimental results show that the novel initialization method can improve the performance of the semi-supervised clustering to a great extent, especially when the labeled samples are short.

## 2 Two Kinds of Initialization Methods for Semi-supervised Clustering

The initialization of semi-supervised clustering is valuable for the further clustering since it can reduce the iterative times of clustering and improve the quality of clusters. The following are two kinds of initialization methods that are widely used in the semi-supervised clustering. They are the random sample method [2] and the modified KKZ initialization method [3].

### 2.1 Random Sample Initialization Method

The random sample method follows a naive way to initialize the seeds of clusters, either using randomly selected input samples, or random parameters non-heuristically generated from the inputs. In the semi-supervised clustering, there are a few labeled samples so that they just randomly label the unlabeled data [2, 7]. The random sample method is shown as follows.

**Random sample method:**

- 1) For  $i=1, \dots, k$ ,  $\mathbf{x} \in S_i$  is the labeled data,  $|S_i|$  is the number of samples whose label is  $i$ . Then the seeds of clusters which have the labeled samples is:
 
$$\mathbf{c}_i = (\sum_{\mathbf{x} \in S_i} \mathbf{x}) / |S_i|$$
- 2) If  $k < K$ ,  $K$  is the number of category, then for  $j=k+1, \dots, K$ ,  $\mathbf{c}_j = \text{random}\{\mathbf{x} | \mathbf{x} \in U \& \mathbf{x} \neq \mathbf{c}_i, \mathbf{c}_i \text{ is the existing centroids}\}$ ,  $U$  is the unlabeled datasets.

**Fig. 1.** Random Sample initialization method



Sometimes, this kind of method can give a good clustering result but it is highly sensitive to the quality of selected seeds.

### 2.2 Modified KKZ Initialization Method

The modified KKZ method utilizes the sorted pairwise distances for initialization which was proposed by Katsavounidis et al. [4, 8]. Shi has modified this method for the semi-supervised clustering [3]. From the experimental results, the modified KKZ is proved to be one of the most effective initialization methods. It is stated as follows.

**The modified KKZ method:**

- 1) If there are no labeled samples, initialize the first seed of cluster using the input with the maximal norm, i.e.  $c_1 = \mathbf{x}_j = \operatorname{argmax} \{ \|\mathbf{x}_j - (\sum \mathbf{x}) / |\mathbf{x}| \| \}$ . Where  $|\mathbf{x}|$  is the number of  $x$ .
- 2) If there are labeled samples, use the labeled data as seeds of clusters as the first step presented in Random sample method.
- 3) If the labels of labeled data are  $1, \dots, k$  and  $k < K$ ,  $K$  is the number of categories, then for  $j = k+1, \dots, K$ , each  $c_j$  is initialized in the following way. For each unlabeled data  $\mathbf{x}_p$ , calculate its distance to the closest seed of cluster  $d_j = \min \{ \|\mathbf{x}_p - \mathbf{c}_i\| : \text{for all existing } c_i \}$ , and set  $c_j = \mathbf{x}_p = \operatorname{argmax}_{\mathbf{x}} \{ d_j \}$

Fig. 2. Modified KKZ initialization method

The seeds of clusters can be well separated by the modified KKZ method. Such initialization method can bring the better and stable results compared with the random sample method.

The above two kinds of initialization methods ignore the unlabeled data which may contain some useful information for clustering. This inspires us to design a new mechanism which can form the seeds of clusters by using the whole information not only from the labeled samples but also from the unlabeled samples.

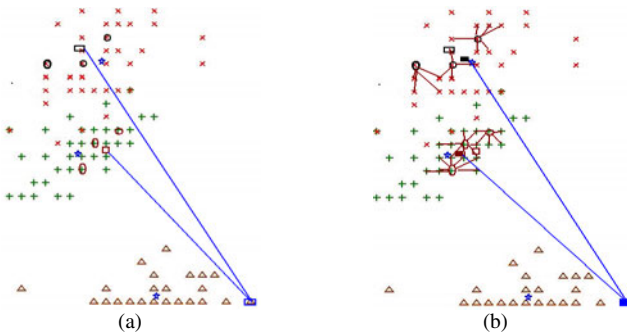
## 3 LDP Based Initialization Method for Semi-supervised Clustering

In this section, we present a novel initialization method called Labeled Data Propagation (LDP) in detail. First, the initial seed selection method is addressed. And then a new similarity measure is presented. At last, the LDP based initialization method is described.

### 3.1 Initial Seed Selection of Clusters

Normally, the incomplete information on the labels results from two situations. One is from the small number of labeled samples and another is from the non-labeled

samples. For the latter situation, the modified KKZ method selects one sample from a cluster as a seed which should be the furthest from the existing seeds (centroids) of the other clusters, as shown in Fig. 3(a) in which the empty rectangle on the bottom is a selected seed but without the label. For the first situation, there are only a few labeled samples being used to form the seeds of clusters. The modified KKZ method just collects the labeled ones to be the seeds. So it loses more information coming from the unlabeled samples. To make full use of both labeled and unlabeled samples to form the seeds, we, in the study, employ the known labels to implement the label propagation to the candidate samples, as shown in Fig. 3(b). The non-labeled neighbors around the given labeled sample are called the candidates which are then examined by the proposed similarity measure and accordingly assigned the correct labels. With more labeled samples to generate the initial seeds, they would be more exact than the seeds from modified KKZ approach. The similarity measure involved is introduced in the following section.



**Fig. 3.** Two seed selection methods (a) The modified KKZ method, (b) Our method

### 3.2 The Similarity Measure

Before calculating the similarity between data samples, some definitions are given at first. Suppose there are a set of  $N$  data samples  $X$  and the corresponding labels  $Y$ , represented by vectors  $\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_N$  and labels  $y_1, y_2, \dots, y_N$  respectively, where  $\{y_1, y_2, \dots, y_N\} \in \{1, 2, \dots, K\}$ ;  $\mathbf{x}$  is the high-dimensional vector and  $K$  is the number of clusters. Let the cluster centroids be  $\mathbf{c}_1, \mathbf{c}_2, \dots, \mathbf{c}_K$ .

In [9], it indicates that the similarity between data samples is not only related to their inherent information but also to their external information, say the number of their common neighbors. From this point of view, a new similarity measure is put forward, which contains two parts. The first part measures the number of common neighbors of two concerned samples and the second part measures the similarity between them by means of the cosine function.

To identify the neighbors of a given sample denoted by  $\mathbf{x}_i$ , a threshold  $\theta$  is set to control the degree of similarity. Suppose  $\mathbf{x}_j$  is the candidate neighbor of  $\mathbf{x}_i$ , if  $\cos(\mathbf{x}_i, \mathbf{x}_j) \geq \theta$ ,  $0 \leq \theta \leq 1$ , then  $\mathbf{x}_j$  should be considered as the neighbor of  $\mathbf{x}_i$ . Once we get all neighbors around the sample  $\mathbf{x}_i$ , the number of common neighbors of  $\mathbf{x}_i$  can be

calculated by Eq. (1). In the same way, the number of common neighbors of centroid  $c_k$  can be obtained from Eq. (2).

$$comn(x_i, x_j) = \sum_{m=1}^n M[i, m] * M[j, m]^T \tag{1}$$

$$comn(c_k, x_j) = \sum_{m=1}^n M[n+k, m] * M[j, m]^T \tag{2}$$

where  $M$  is an  $N \times (N+K)$  adjacency matrix which includes  $N$  data samples and  $K$  centroids of clusters.  $M[i, j]$  is 1 or 0 depending on whether  $x_j$  is the neighbor of  $x_i$  or not.

Therefore, the new similarity measure can be given as following:

$$sim(c_k, x_j) = \alpha * cnd(c_k, x_j) + (1 - \alpha) * \cos(c_k, x_j) \tag{3}$$

$$cnd(c_k, x_j) = \frac{comn(c_k, x_j)}{C \max_k} \tag{4}$$

where  $C \max_k$  is the largest possible value of  $comn(c_k, x_j)$ , and  $\alpha$  is the coefficient set by the user.

Similarly, we give the similarity measure between any pair of data samples which is defined as follows.

$$sim(x_i, x_j) = \alpha * cnd(x_i, x_j) + (1 - \alpha) * \cos(x_i, x_j) \tag{5}$$

$$cnd(x_i, x_j) = \frac{comn(x_i, x_j)}{L \max_i} \tag{6}$$

where  $L \max_i = \max_i(comn(x_i, x_j))$ .

### 3.3 LDP Initialization Method

The LDP method finds the candidates to be propagated with labels from two aspects. From a labeled sample to an unlabeled sample, the unlabeled one could become a candidate if it holds the maximal similarity value for the given labeled sample. Reversely from the current candidate to all samples, the maximal similarity value between them can be worked out. Then, a ratio of the above two different similarity values is introduced in the study which is used as a criterion for the label propagation. Let  $R$  denote the ratio, and  $\tau$  denote a threshold which is a tuneable parameter. The ratio can be calculated by

$$R = \frac{\max(sim(x_i^{(y)}, x_j))}{\max(sim(x_j, x_p))} \tag{7}$$

where  $x_i^{(y)}$  is the labeled sample,  $x_j$  is the unlabeled sample and  $x_p$  is the other samples in the dataset. If  $R \geq \tau$ , the concerned candidate sample  $x_j$  can be added the same

label as  $x_i^{(y)}$ . By sorting all ratios in the descending order, the top  $t$  candidates can be found out, where  $t$  is predefined by the user.

The above is the propagation process for unlabeled samples. In the extreme situation, there is no any labeled data in a cluster. In this case, the sample in the current cluster could be a seed if it has the minimal similarity value to the centroid of the other cluster. The minimal similarity between the unlabeled and the known centroid is given in expression (8).

$$\arg \min \{ \text{sim}(c, x_j), \text{for } x_j \neq c, x_j \in U \} \tag{8}$$

where  $c$  is the existing centroid, and  $U$  is the set of unlabeled data samples.

By using the LDP based initialization method, the seeds obtained from the unlabeled samples are well separated. The detailed steps of this initialization method are given below.

**LDP based initialization method:**

- 1) From the labeled samples, it calculates the similarity between labeled samples and unlabeled ones:  $\text{sim}(x_i, x_j)$ . There are  $t$  unlabeled data as candidates which are the elements in the top  $t$  maximal  $\text{sim}(x_i, x_j)$ . Then from each candidate  $x_j$ , it calculates  $\max(\text{sim}(x_j, x_p))$ . The  $R$  is given by
 
$$R = \frac{\text{sim}(x_i^{(y)}, x_j)}{\max(\text{sim}(x_j, x_p))}$$
 If  $R$  is large enough, the concerned candidate sample  $x_j$  can be added the same label as  $x_i^{(y)}$ .
- 2) If the labels of labeled samples are  $1, \dots, k$  and  $k < K$ .  $K$  is the number of category. Then  $k+1, \dots, K$ , each centroid  $c_j$  is initialized in the following way:
 
$$c_j = x_j = \arg \min \{ \text{sim}(c, x_j), \text{for } x_j \neq c, x_j \in U \}.$$
 where  $c_j$  is the existing centroid, and  $U$  is the set of unlabeled data samples.

Fig. 4. LDP based initialization method

## 4 Experiments

### 4.1 Dataset

The experiment utilizes two document datasets which are Classic400 and tr11 respectively. Classic400 is a sub-dataset from the classic document set which includes four categories CACM, CISI, CRANFIELD, and MEDLINE. tr11 is derived from Text Retrieval Conference (TREC) collections. These two datasets have been preprocessed with different toolkit by many researchers.

The selected datasets provide a good representation of different characteristics: the number of categories is from 3 to 9; the *Balance*, which is the specific value, is from 0.0455 to 0.5.

**Table 1.** Statistics of text datasets (  $N$  is the total number of documents,  $|V|$  is the total number of words,  $K$  is the number of category,  $max\_cat\_size$  is the maximum number of documents,  $min\_cat\_size$  is the minimum number of documents , *Balance* is the ratio of  $min\_cat\_size$  to  $max\_cat\_size$  )

Data	$N$	$K$	$max\_class\_size$	$min\_class\_size$	$ V $	<i>Balance</i>
Classic400	400	3	200	100	6205	0.5
tr11	414	9	132	6	6424	0.0455

## 4.2 Experiment Setup

In the experiment, two kinds of semi-supervised clustering methods are involved: discriminative (or distance/ similarity-based) approach and generative (or model-based) approach [7,10]. The discriminative approach, such as clustering based on graph, determines a distance or similarity function between pairs of data samples and then groups the similar samples together into clusters. The generative approach, on the other hand, attempts to learn generative models from the data, with each model representing one particular cluster. In the discriminative approach, for high-dimensional textual data, similarity based measures are commonly used. Compared to similarity-based methods, the generative method offers the better interpretability since the resulting model for each cluster can directly characterize the concerned cluster. Model-based partition clustering algorithm often has a better smooth clustering to avoid the local optimization. In the paper, we employ Seeded Semi-supervised K-means Clustering algorithm (SSKC) [2] corresponding to the discriminative approach and Probability EM Semi-supervised Clustering algorithm (PECSC) [6, 7, 11] corresponding to the generative approach to do the experiments.

For the high dimensional document datasets, we set the threshold  $\theta=0.03$ , the thresholds  $\alpha=0.85$  and  $\beta=0.75$ . When the number of labeled samples is small, we set  $t=5$ .

Considering the difference between semi-supervised clustering and semi-supervised classification, we give the following two different cases when the datasets are labeled data.

- Complete labels. We randomly pick put the labeled samples from the whole dataset with different sampling probabilities which are 1.5%, 2.5%, 5.0%, 7.5%, 10.0%, 20.0%, 30.0%, and 40.0% respectively.

- Incomplete labels. We first randomly choose the half of all categories as candidates. Then we select some labeled samples from the candidates with the probabilities of 1.5%, 2.5%, 5.0%, 7.5%, 10.0%, 20.0%, 30.0%, and 40.0% respectively.

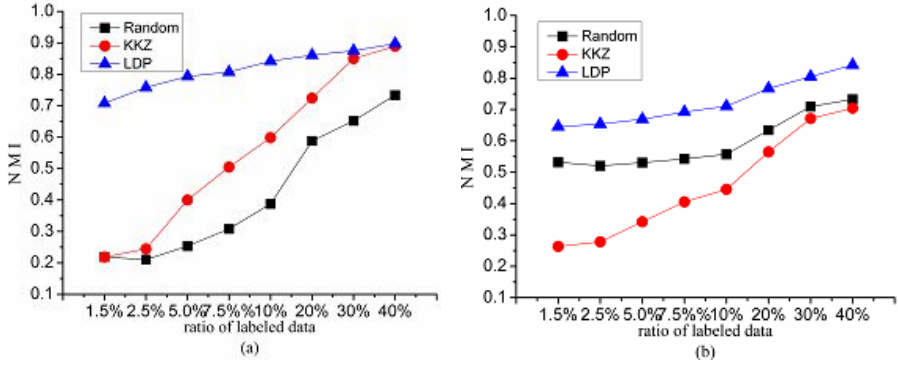


Fig. 5. Comparisons of NMI among three initialization methods with different datasets based on PECS (the completed label situation). (a) Classic400, (b) tr11.

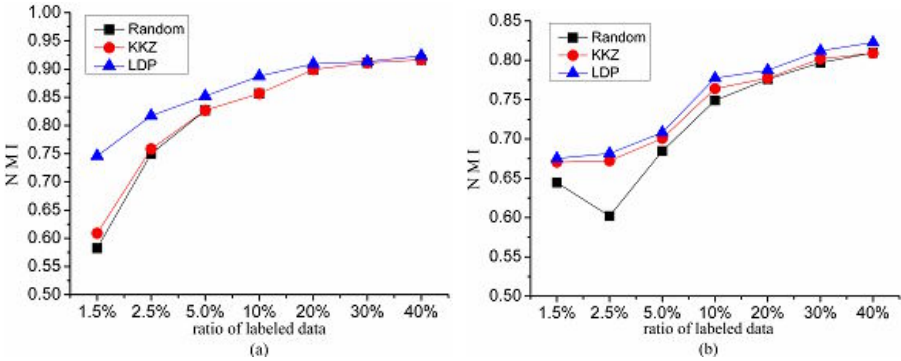
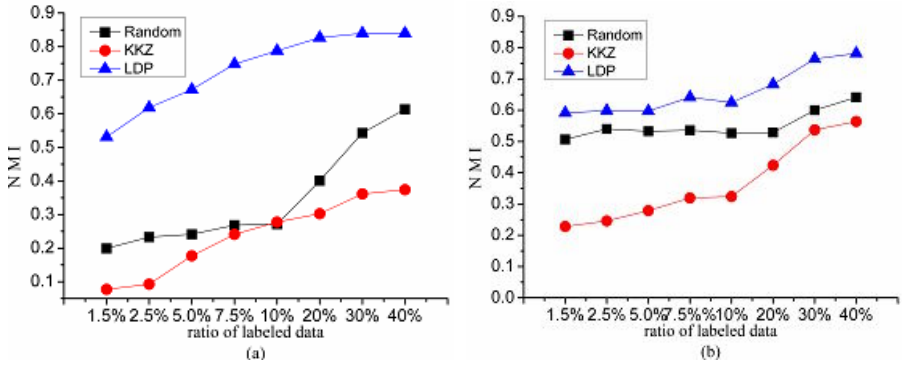


Fig. 6. Comparisons of NMI among three initialization methods with different datasets based on SSKC (the completed label situation). (a) Classic400, (b) tr11.

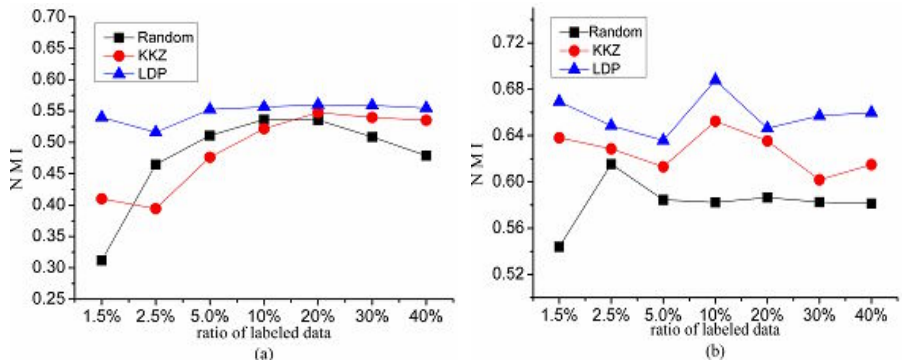
We use the normalized mutual information (NMI) to calculate the accuracy of semi-supervised clustering [3]. Because the labeled samples which are selected at random will have a great impact on the clustering results, for each algorithm and each percentage setting, we repeat the random sampling process ten times and report the average of NMI values (i.e. *Mean*) for the clustering results. The *Max* is the maximal NMI value among ten runs and the *Min* is the minimal NMI value. We use  $F = |Max - Min|/Mean$  to estimate the changes of clustering results.

#### 4.4 Experimental Results and Discussions

The initialization method based on two semi-supervised clustering algorithms, which are the basic semi-supervised clustering algorithms. From the experimental results of the initialization methods, as shown in Figs 5, 6, 7 and 8, we can find that the proposed initialization method performs very well on the selected document datasets. However, the experiments on different document datasets also show that the LDP method could not always make the clustering accurately. If the category  $k$  is big, especially if the discrepancy is not enough, the performance of LDP is not very good.



**Fig. 7.** Comparisons of NMI among three initialization methods with different datasets based on PECS (the incomplete label situation). (a) Classic400, (b) tr11.



**Fig. 8.** Comparisons of NMI among three initialization methods with different datasets based on SSK (the incomplete label situation). (a) Classic400 and (b) tr11.

In the case of the completed labels (Figs 5 and 6), when the percentages of labeled data to all data samples are 1.5%, 2.5% and 5% respectively, the clustering results with LDP initialization are the best comparing with the other initialization methods. In addition, the NMI increases fast as the percentage of labeled data goes up. But when the *Balance* is small or the number of category is big (tr11 in Figs 5 and 6), there will be no labeled samples or short labeled samples in some categories. This is because the data number is small in these categories. As a result, it is hard to find the optimal seeds of clusters so that the initialization has a limited improvement. In the case of the incomplete labels (Figs 7 and 8), when the percentages of labeled data to all data samples are 1.5%, 2.5% and 5% respectively, the LDP initialization performs best among three initialization methods.

The NMI has great improvement when the ratio of labeled data is small. We show the clustering results in Tables 2 and 3 when the ratio of labeled sample is 10%. And we use  $F$  to estimate the changes of clustering results with different initialization methods in some datasets. The changes of clustering results with LDP initialization method keep a low level. With the same dataset, both the Min and Max of the

clustering results with LDP initialization method is the biggest. The *Balance* is close to 1, the effect of the new method is better.

**Table 2.** Comparisons for the Stability of NMI (The labeled data is 10% and the labels are complete, Min is the minimal NMI value of clustering results, Max is the maximal NMI value of clustering results, Mean is the mean NMI value of clustering results, and F is the fluctuation of clustering results)

Data	Initialization method	Min	Max	Mean	F
Classic400	Random	0.3103	0.5020	0.3876	0.4946
	KKZ	0.4946	0.6920	0.5983	0.3299
	LDP	0.7802	0.8926	0.8417	0.1335
tr11	Random	0.2063	0.6160	0.5573	0.7352
	KKZ	0.3890	0.5146	0.4454	0.2820
	LDP	0.6598	0.7492	0.7107	0.1258

**Table 3.** Comparisons for the Stability of NMI (The labeled data is 10% and the labels are incomplete, Min is the minimal NMI value of clustering results, Max is the maximal NMI value of clustering results, Mean is the mean NMI value of clustering results, and F is the fluctuation of clustering results)

Data	Initialization method	Min	Max	Mean	F
Classic400	Random	0.1724	0.3792	0.2714	0.7251
	KKZ	0.2055	0.3496	0.2779	0.5185
	LDP	0.6687	0.8539	0.7877	0.2351
tr11	Random	0.3700	0.6222	0.5267	0.4788
	KKZ	0.2462	0.4185	0.3240	0.5318
	LDP	0.5675	0.6694	0.6247	0.1631

From the above three tables, we find that three main factors would influence the performance of the initialization method.

(i) The completed labels can make the initialization more effectively for semi-supervised clustering. However, in the real world, there are so many incomplete label situations, which is imperative to enhance initialization with incomplete labels. The random initialization method is uncontrollable to create the seeds of clusters so that the effect of the labeled samples is very limited. The modified KKZ initialization method just uses the labeled samples to treat the cluster centroids as the seeds. The LDP based initialization method uses a new similarity measure which is more powerful to find the representative samples as the cluster centroids. If the datasets are not well separated, the similarity only based on the cosine function would perform badly in the initialization of semi-supervised clustering algorithm. As the new similarity measure criterion introduces more information for seed selection, it is



therefore more reasonable and effective in the initialization of semi-supervised clustering algorithm.

(ii) The number of labeled samples is very important for clustering. The figures show that the low labeled sample number will lead up to the low NMI value which implies the low clustering accuracy. So we use the label propagation approach to increase the labeled sample number, as a result the seeds of clusters are more accurate.

(iii) The density of data samples also has an impact on the clustering results. We propagate labels with a new method. From labeled samples to find candidates, from the candidates to get the maximal similarity value, we use a ratio of the two different similarity values decide to propagate labels to candidates. This method make the initialization perform better in the situation that density of categories is different. Further more, when a labeled sample is distant from any other data sample, we also could use it to propagate label to improve the seeds of clusters.

## 5 Conclusion

In this paper, we propose a labeled data propagation (LDP) initialization method for semi-supervised clustering problems. Comparing with two existing initialization methods based on two semi-supervised clustering models, the new method effectively improves the accuracy of clustering. Especially, when the ratio of labeled data is small, the accuracy of clustering results with proposed initialization method is the best.

The experimental results show that the LDP based initialization method is very effective for the semi-supervised clustering. The LDP method uses the label propagation approach to increase the number of labeled samples. And the new similarity measure is more effective to find the representative samples as cluster seeds by which the better groups of the datasets can be obtained in the initial step of algorithm. What's more, the importance of initialization shown in the experiments suggests that we should do further work to mine the labeled samples to enhance the semi-supervised clustering algorithm.

**Acknowledgments.** This work has been partly supported by the National High Technology Research and Development Program of China (No.2008AA04Z107) and the Natural Science Foundation of China under Grant Nos. 70771019 and 70871016.

## References

1. Zhou, Z.H., Zhang, D.C., Yang, Q.: Semi-Supervised Learning with Very Few Labeled Training Examples. In: 22nd AAAI Conference on Artificial Intelligence, pp. 675–680. AAAI Press, Vancouver (2007)
2. Basu, S., Banerjee, A., Mooney, R.: Semi-supervised Clustering by Seeding. In: 19th International Conference Machine Learning, pp. 19–26. Morgan Kaufmann Press, Sydney (2002)
3. Zhong, S.: Semi-supervised Model-based Document Clustering: A Comparative Study. *J. Mach. Learn.* 65, 3–29 (2006)
4. Katsavounidis, I., Kuo, C., Zhang, Z.: A New Initialization Technique for Generalized Lloyd Iteration. *J. Sig. Proc. Lett.* 1, 144–146 (1994)

5. Sun, X., Li, K.L., Zhao, R.: Global optimization for semi-supervised K-means. In: Asia-Pacific Conference on Information Processing, pp. 410–413. IEEE Press, Shen Zhen (2009)
6. Zhu, X.J., Ghahramani, Z.: Learning from Labeled and Unlabeled Data with Label Propagation. Technical Report CMU-CALD-02-107, Carnegie Mellon Univ. (2002)
7. Zhong, S., Ghosh, J.: A unified framework for model-based clustering. *J. Mach. Learn. Resear.* 4, 1001–1037 (2003)
8. He, J., Lan, M., Tan, C.L., Sung, S.Y., Low, H.B.: Initialization of Cluster Refinement Algorithms: A Review and Comparative Study. In: IEEE International Joint Conference Neural Networks, pp. 297–302. IEEE Press, Budapest (2004)
9. Luo, C., Li, Y.J., Chung, S.M.: Text Document Clustering Based on Neighbors. *J. Data & Kno. Engin.* 68, 1271–1288 (2009)
10. Nigam, K., Mccallum, A., Thrun, S., Mitchell, T.: Text classification from labeled and unlabeled documents using EM. *J. Mach. Learn.* 39, 103–134 (2000)
11. Nigam, K.: Using Unlabeled Data to Improve Text Classification. Doctoral Dissertation, School of Computer Science, Carnegie Mellon University (2001)

# Constructing and Mapping Fuzzy Thematic Clusters to Higher Ranks in a Taxonomy

Boris Mirkin<sup>1,2</sup>, Susana Nascimento<sup>3</sup>, Trevor Fenner<sup>1</sup>, and Luís Moniz Pereira<sup>3</sup>

<sup>1</sup> School of Computer Science, Birkbeck University of London, London, WC1E 7HX, UK

<sup>2</sup> Division of Applied Mathematics, Higher School of Economics, Moscow, RF

<sup>3</sup> Computer Science Department and Centre for Artificial Intelligence (CENTRIA), Faculdade de Ciências e Tecnologia, Universidade Nova de Lisboa, Caparica, Portugal

**Abstract.** We present a novel methodology for mapping a system such as a research department to a related taxonomy in a thematically consistent way. The components of the structure are supplied with fuzzy membership profiles over the taxonomy. Our method generalizes the profiles in two steps: first, by fuzzy clustering, and then by mapping the clusters to higher ranks of the taxonomy. To be specific, we concentrate on the Computer Sciences area represented by the taxonomy of ACM Computing Classification System (ACM-CCS). We build fuzzy clusters of the taxonomy leaves according to the similarity between individual profiles by using a novel, additive spectral, fuzzy clustering method that, in contrast to other methods, involves a number of model-based stopping conditions. The clusters are not necessarily consistent with the taxonomy. This is formalized by a novel method for parsimoniously elevating them to higher ranks of the taxonomy using an original recursive algorithm for minimizing a penalty function that involves “head subjects” on the higher ranks of the taxonomy along with their “gaps” and “offshoots”. An example is given illustrating the method applied to real-world data.

## 1 Introduction

The last decade has witnessed an unprecedented rise of the concept of ontology as a computationally feasible tool for knowledge maintenance. For example, the usage of Gene Ontology [6] for interpretation and annotation of various gene sets and gene expression data is becoming a matter of routine in bioinformatics (see, for example, [14] and references therein).

The goal of this paper is to develop a framework for representation of the activities of an organization or any other system under consideration, in terms of a taxonomy. We first build profiles for its constituent entities in terms of the taxonomy and then thematically generalize them to higher ranks of the taxonomy.

To represent a functioning structure over a taxonomy is to indicate those topics in the taxonomy that most fully express the structure’s working in its relation to the taxonomy. To make the representation thematically consistent and parsimonious, we have developed a two-phase generalization approach. The first phase generalizes over the structure by building clusters of taxonomy topics according to the functioning of the system. The second phase takes the clusters as query sets in the taxonomy and parsimoniously maps

them to higher ranks of the taxonomy. Both entity profiles and thematic clusters derived at the first phase are fuzzy in order to better reflect the real world objects, so that the elevating method applies to fuzzy clusters. It should be pointed out that both building fuzzy profiles and finding fuzzy clusters are research activities well documented in the literature; yet the issues involved in this project led us to develop original schemes of our own including an efficient method for fuzzy clustering combining the approaches of spectral and approximation clustering [12].

We apply these constructions in two areas: (i) to visualize activities of Computer Science research organizations; and (ii) to discern the complexes of mathematical ideas according to classes taught in regular teaching courses in a university department. We take the popular ACM Computing Classification System (ACM-CCS), a conceptual four-level classification of the Computer Science subject area as a pre-specified taxonomy for (i), and the three-layer Mathematics Subject Classification MSC2010 developed by the Mathematical Reviews and Zentralblatt Mathematics editors (see <http://www.ams.org/mathscinet/msc/msc2010.html>), for (ii). In what follows the focus is mainly on the application (i) to research organizations. The paper is organized according to the structure of our approach: Section 2 describes an e-system we developed for getting ACM-CCS leaves fuzzy membership profiles from Computer Science researchers, Section 3 describes our method for deriving fuzzy clusters from the profiles, and Section 4 presents our parsimonious elevation method to generalize to higher ranks in a taxonomy tree.

## 2 Taxonomy-Based Profiles

### 2.1 Representing over the ACM-CCS Taxonomy

In the case of investigation of activities of a university department or center, a research team's profile can be defined as a fuzzy membership function on the set of leaf-nodes of the taxonomy under consideration so that the memberships reflect the extent of the team's effort put into corresponding research topics.

In this case, the ACM Computing Classification System (ACM-CCS) [1] is used as the taxonomy. ACM-CCS comprises eleven major partitions (first-level subjects) such as *B. Hardware*, *D. Software*, *E. Data*, *G. Mathematics of Computing*, *H. Information Systems*, etc. These are subdivided into 81 second-level subjects. For example, item *I. Computing Methodologies* consists of eight subjects including *I.1 SYMBOLIC AND ALGEBRAIC MANIPULATION*, *I.2 ARTIFICIAL INTELLIGENCE*, *I.5 PATTERN RECOGNITION*, etc. They are further subdivided into third-layer topics as, for instance, *I.5 PATTERN RECOGNITION* which is represented by seven topics including *I.5.3 Clustering*, *I.5.4 Applications*, etc.

Taxonomy structures such as the ACM-CCS are used, mainly, as devices for annotation and search for documents or publications in collections such as that on the ACM portal [1]. The ACM-CCS tree has been applied also as: a gold standard for ontologies derived by web mining systems such as the CORDER engine [17]; a device for determining the semantic similarity in information retrieval [9] and e-learning

applications [18,5]; and a device for matching software practitioners’ needs and software researchers’ activities [4].

Here we concentrate on a different application of ACM-CCS – a generalized representation of a Computer Science research organization that can be used for overviewing scientific subjects that are being developed in the organization, assessing the scientific issues in which the character of activities in organizations does not fit well onto the classification – these can potentially be the growth points, and help with planning the restructuring of research and investment.

### 2.2 E-Screen Survey Tool

Fuzzy profiles are derived from either automatic analysis of documents posted on the web by the teams or by explicitly surveying the members of the department. The latter option is especially convenient in situations in which the web contents do not properly reflect the developments, for example, in non-English speaking countries with relatively underdeveloped internet infrastructures for the maintenance of research results. We developed an interactive survey tool that provides two types of functionality: i) collection of data about ACM-CCS based research profiles of individual members; ii) statistical analysis and visualization of the data and results of the survey on the level of a department. The respondent is asked to select up to six topics among the leaf nodes of the ACM-CCS tree and assign each with a percentage expressing the proportion of the topic in the total of the respondent’s research activity for, say, the past four years. Figure 1 shows a screenshot of the baseline interface for a respondent who has chosen six ACM-CCS topics during her survey session.

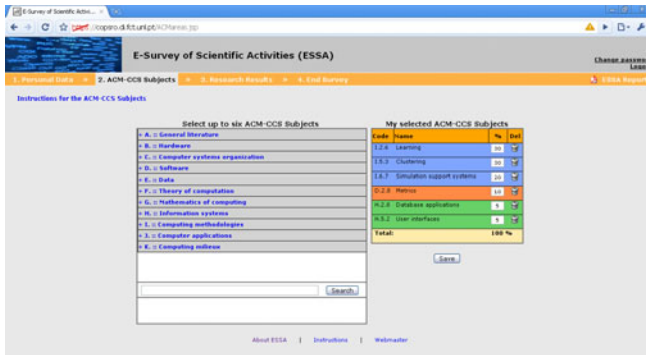


Fig. 1. Screenshot of the interface survey tool for selection of ACM-CCS topics

The set of profiles supplied by respondents forms an  $N \times M$  matrix  $F$  where  $N$  is the number of ACM-CCS topics involved in the profiles and  $M$  the number of respondents. Each column of  $F$  is a fuzzy membership function, rather sharply delineated because only six topics may have positive memberships in each of the columns.

### 3 Representing Research Organization by Fuzzy Clusters of ACM-CCS Topics

#### 3.1 Deriving Similarity between ACM-CCS Research Topics

We represent a research organization by clusters of ACM-CCS topics to reflect thematic communalities between activities of members or teams working on these topics. The clusters are found by analyzing similarities between topics according to their appearances in the profiles. The more profiles contain a pair of topics  $i$  and  $j$  and the greater the memberships of these topics, the greater is the similarity score for the pair.

Consider a set of  $V$  individuals ( $v = 1, 2, \dots, V$ ), engaged in research over some topics  $t \in T$  where  $T$  is a pre-specified set of scientific subjects. The level of research effort by individual  $v$  in developing topic  $t$  is evaluated by the membership  $f_{tv}$  in profile  $f_v$  ( $v = 1, 2, \dots, V$ ).

Then the similarity  $w_{tt'}$  between topics  $t$  and  $t'$  is defined as

$$w_{tt'} = \sum_{v=1}^V \frac{n_v}{n_{max}} f_{tv} f_{t'v}, \quad (1)$$

where the ratios are introduced to balance the scores of individuals bearing different numbers of topics.

To make the cluster structure in the similarity matrix sharper, we apply the spectral clustering approach to pre-process the similarity matrix  $W$  using the so-called Laplacian transformation [8]. First, an  $N \times N$  diagonal matrix  $D$  is defined, with  $(t, t)$  entry equal to  $d_t = \sum_{t' \in T} w_{tt'}$ , the sum of  $t$ 's row of  $W$ . Then unnormalized Laplacian and normalized Laplacian are defined by equations  $L = D - W$  and  $L_n = D^{-1/2} L D^{-1/2}$ , respectively. Both matrices are semipositive definite and have zero as the minimum eigenvalue. The minimum non-zero eigenvalues and corresponding eigenvectors of the Laplacian matrices are utilized then as relaxations of combinatorial partition problems [16,8]. Of comparative properties of these two normalizations, the normalized Laplacian, in general, is considered superior [8]. Since the additive clustering approach described in the next section relies on maximum rather than minimum eigenvalues, we use the Laplacian pseudoinverse transformation, Lapin for short, defined by

$$L_n^-(W) = \tilde{Z} \tilde{A}^{-1} \tilde{Z}'$$

where  $\tilde{A}$  and  $\tilde{Z}$  are defined by the spectral decomposition  $L_n = ZAZ'$  of matrix  $L_n = D^{-1/2}(D - W)D^{-1/2}$ . To specify these matrices, first, set  $T'$  of indices of elements corresponding to non-zero elements of  $A$  is determined, after which the matrices are taken as  $\tilde{A} = A(T', T')$  and  $\tilde{Z} = Z(:, T')$ . The choice of the Lapin transformation can be explained by the fact that it leaves the eigenvectors of  $L_n$  unchanged while inverting the non-zero eigenvalues  $\lambda \neq 0$  to those  $1/\lambda$  of  $L_n^-$ . Then the maximum eigenvalue of  $L_n^-$  is the inverse of the minimum non-zero eigenvalue  $\lambda_1$  of  $L_n$ , corresponding to the same eigenvector.

### 3.2 Additive-Spectral Fuzzy Clustering

In spite of the fact that many fuzzy clustering algorithms have been developed already [2], [7], most of them are ad hoc and, moreover, they all involve manually specified parameters such as the number of clusters or threshold of similarity without providing any guidance for choosing the. We apply a model-based approach of additive clustering, combined with the spectral clustering approach, to develop a novel fuzzy clustering method that is both practical and supplied with model-based parameters helping to choose the right number of clusters.

Thematic similarities  $a_{tt'}$  between topics are but manifested expressions of some hidden patterns within the organization which can be represented by fuzzy clusters in exactly the same manner as the manifested scores in the definition of the similarity  $w_{tt'}$  (1). We propose to formalize a thematic fuzzy cluster as represented by two items: (i) a membership vector  $u = (u_t)$ ,  $t \in T$ , such that  $0 \leq u_t \leq 1$  for all  $t \in T$ , and (ii) an intensity  $\mu > 0$  that expresses the extent of significance of the pattern corresponding to the cluster, within the organization under consideration. With the introduction of the intensity, applied as a scaling factor to  $u$ , it is the product  $\mu u$  that is a solution rather than its individual co-factors. Given a value of the product  $\mu u_t$ , it is impossible to tell which part of it is  $\mu$  and which  $u_t$ . To resolve this, we follow a conventional scheme: let us constrain the scale of the membership vector  $u$  on a constant level, for example, by a condition such as  $\sum_t u_t = 1$  or  $\sum_t u_t^2 = 1$ , then the remaining factor will define the value of  $\mu$ . The latter normalization better suits the criterion implied by our fuzzy clustering method and, thus, is accepted further on.

Our additive fuzzy clustering model follows that of [15,10,13] and involves  $K$  fuzzy clusters that reproduce the pseudo-inverted Laplacian similarities  $a_{tt'}$  up to additive errors according to the following equations:

$$a_{tt'} = \sum_{k=1}^K \mu_k^2 u_{kt} u_{kt'} + e_{tt'}, \tag{2}$$

where  $u_k = (u_{kt})$  is the membership vector of cluster  $k$ , and  $\mu_k$  its intensity.

The item  $\mu_k^2 u_{kt} u_{kt'}$  expresses the contribution of cluster  $k$  to the similarity  $a_{tt'}$  between topics  $t$  and  $t'$ , which depends on both the cluster's intensity and the membership values. The value  $\mu^2$  summarizes the contribution of intensity and will be referred to as the cluster's weight.

To fit the model in (2), we apply the least-squares approach, thus minimizing the sum of all  $e_{tt'}^2$ . Since  $A$  is definite semi-positive, its first  $K$  eigenvalues and corresponding eigenvectors form a solution to this if no constraints on vectors  $u_k$  are imposed. Additionally, we apply the one-by-one principal component analysis strategy for finding one cluster at a time this makes the computation feasible and is crucial for determining the number of clusters. Specifically, at each step, we consider the problem of minimization of a reduced to one fuzzy cluster least-squares criterion

$$E = \sum_{t,t' \in T} (b_{tt'} - \xi u_t u_{t'})^2 \tag{3}$$

with respect to unknown positive  $\xi$  weight (so that the intensity  $\mu$  is the square root of  $\xi$ ) and fuzzy membership vector  $u = (u_t)$ , given similarity matrix  $B = (b_{tt'})$ .

At the first step,  $B$  is taken to be equal to  $A$ . Each found cluster changes  $B$  by subtracting the contribution of the found cluster (which is additive according to model (2)), so that the residual similarity matrix for obtaining the next cluster will be  $B - \mu^2 uu^T$  where  $\mu$  and  $u$  are the intensity and membership vector of the found cluster. In this way,  $A$  indeed is additively decomposed according to formula (2) and the number of clusters  $K$  can be determined in the process.

Let us specify an arbitrary membership vector  $u$  and find the value of  $\xi$  minimizing criterion (3) at this  $u$  by using the first-order condition of optimality:

$$\xi = \frac{\sum_{t,t' \in T} b_{tt'} u_t u_{t'}}{\sum_{t \in T} u_t^2 \sum_{t' \in T} u_{t'}^2},$$

so that the optimal  $\xi$  is

$$\xi = \frac{\mathbf{u}' B \mathbf{u}}{(\mathbf{u}' \mathbf{u})^2} \tag{4}$$

which is obviously non-negative if  $B$  is semi-positive definite.

By putting this  $\xi$  in equation (3), we arrive at

$$E = \sum_{t,t' \in T} b_{tt'}^2 - \xi^2 \sum_{t \in T} u_t^2 \sum_{t' \in T} u_{t'}^2 = S(B) - \xi^2 (\mathbf{u}' \mathbf{u})^2,$$

where  $S(B) = \sum_{t,t' \in T} b_{tt'}^2$  is the similarity data scatter.

Let us denote the last item by

$$G(u) = \xi^2 (\mathbf{u}' \mathbf{u})^2 = \left( \frac{\mathbf{u}' B \mathbf{u}}{\mathbf{u}' \mathbf{u}} \right)^2, \tag{5}$$

so that the similarity data scatter is the sum:

$$S(B) = G(u) + E \tag{6}$$

of two parts,  $G(u)$ , which is explained by cluster  $(\mu, u)$ , and  $E$ , which remains unexplained.

An optimal cluster, according to (6), is to maximize the explained part  $G(u)$  in (5) or its square root

$$g(u) = \xi \mathbf{u}' \mathbf{u} = \frac{\mathbf{u}' B \mathbf{u}}{\mathbf{u}' \mathbf{u}}, \tag{7}$$

which is the celebrated Rayleigh-Ritz quotient, whose maximum value is the maximum eigenvalue of matrix  $B$ , which is reached at its corresponding eigenvector, in the unconstrained problem.

This shows that the spectral clustering approach is appropriate for our problem. According to this approach, one should find the maximum eigenvalue  $\lambda$  and corresponding normed eigenvector  $z$  for  $B$ ,  $[\lambda, z] = \Lambda(B)$ , and take its projection to the set of admissible fuzzy membership vectors.



Our clustering approach involves a number of model-based criteria for halting the process of sequential extraction of fuzzy clusters:

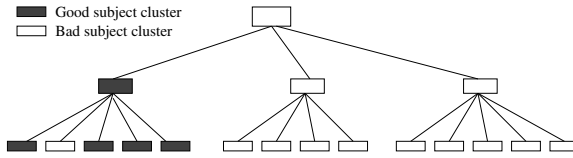
1. The optimal value of  $\xi$  (4) for the spectral fuzzy cluster becomes negative.
2. The contribution of a single extracted cluster becomes too low, less than a pre-specified  $\tau > 0$  value.
3. The residual scatter  $E$  becomes smaller than a pre-specified  $\epsilon$  value, say less than 5% of the original similarity data scatter.

The described one-by-one fuzzy additive-spectral thematic cluster extraction algorithm is referred to as the FADDI-S. It combines three different approaches: additive clustering [15,10,13], spectral clustering [16,8,20], and relational fuzzy clustering [2,3] and adds an edge to each. In the context of additive clustering, fuzzy approaches were considered only by [13], yet in a very restricted setting: (a) the clusters intensities are assumed constant there, (b) the number of clusters is pre-specified, and (c) the fitting method is very local and computationally intensive - these all restrictions are overcome in FADDI-S. The spectral clustering approach is overtly heuristic, whereas FADDI-S is model-based. The criteria used in relational fuzzy clustering are ad hoc whereas that of FADDI-S is model-based, and, moreover, its combined belongingness function values  $\mu u$  are not constrained by the unity as is the case in relational clustering, but rather follow the scales of the relation under investigation, which is in line with the original approach by L. Zadeh [19]. We also carried out experiments to compare the effectiveness of FADDI-S with other popular methods. For example, in recent experiments, [3] compared several most popular relational fuzzy clustering approaches and showed that combining the popular fuzzy c-means approach with an initialization routine was superior to the others; yet FADDI-S outperformed that on the data generated according to the recipe from [3] (details and more experiments are described in [12]).

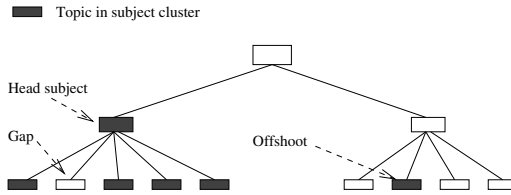
## 4 Parsimonious Elevating Method

To generalize the contents of a thematic cluster, we propose a method for elevating it to higher ranks of the taxonomy so that if all or almost all children of a node in an upper layer belong to the cluster, then the node itself is taken to represent the cluster at this higher level of the ACM-CCS taxonomy (see Fig. 2). Depending on the extent of inconsistency between the cluster and the taxonomy, such elevation can be done differently, leading to different portrayals of the cluster on ACM-CCS tree depending on the relative weights of the events taken into account. A major event is the so-called “head subject”, a taxonomy node covering (some of) leaves belonging to the cluster, so that the cluster is represented by a set of head subjects. The penalty of the representation to be minimized is proportional to the number of head subjects so that the smaller that number the better. Yet the head subjects cannot be elevated too high in the tree because of the penalties for associated events, the cluster “gaps” and “offshoots” the number of them depends on the extent of inconsistency of the cluster versus the taxonomy.

The gaps are head subject’s children topics that are not included in the cluster. An offshoot is a taxonomy leaf node that is a head subject (not elevated). It is not difficult to see that the gaps and offshoots are determined by the head subjects specified in an elevation (see Fig. 3).



**Fig. 2.** Two clusters of second-layer topics, presented with checked and diagonal-lined boxes, respectively. The checked box cluster fits within one first-level category (with one gap only), whereas the diagonal line box cluster is dispersed among two categories on the right. The former fits the classification well; the latter does not.



**Fig. 3.** Three types of features in mapping of a subject cluster to the taxonomy

The total count of head subjects, gaps and offshoots, each weighted by both the penalties and leaf memberships, is used for scoring the extent of the cluster misfit needed for elevating a grouping of research topics over the classification tree. The smaller the score, the more parsimonious the elevation and the better the fit. Depending on the relative weighting of gaps, offshoots and multiple head subjects, different elevations can minimize the total misfit, as illustrated on Fig. 5 later.

Altogether, the set of topic clusters together with their optimal head subjects, offshoots and gaps constitute a parsimonious representation of the organization. Such a representation can be easily accessed and expressed. It can be further elaborated by highlighting those subjects in which members of the organization have been especially successful (i.e., publication in best journals or awards) or distinguished by a special feature (i.e., industrial use or inclusion in a teaching program). Multiple head subjects and offshoots, when they persist at subject clusters in different organizations, may show some tendencies in the development of the science, that the classification has not taken into account yet.

We have proved that a parsimonious lift of a subject cluster can be achieved by recursively building a parsimonious representation for each node of the ACM-CCS tree based on parsimonious representations for its children. In this, we assume that any head subject is automatically present at each of the nodes it covers, unless they are gaps (as presented on Fig. 3). Our algorithm is set as a recursive procedure over the tree starting at leaf nodes.

The procedure determines, at each node of the tree, sets of head gain and gap events to iteratively raise them to those of the parents, under each of two different assumptions that specify the situation at the parental node. One assumption is that the head subject has been inherited at the parental node from its own parent, and the second

assumption is that it has not been inherited but gained in the node only. In the latter case the parental node is labeled as a head subject. Consider the parent-children system as shown in Fig. 4, with each node assigned with sets of gap and head gain events under the above two inheritance of head subject assumptions.

Let us denote the total penalty, to be minimized, under the inheritance and non-inheritance assumptions by  $p_i$  and  $p_n$ , respectively. An elevation result at a given node is defined by a pair of sets (H, G), representing the tree nodes at which events of head gains and gaps, respectively, have occurred in the subtree rooted at the node. We use  $(H_i, G_i)$  and  $(H_n, G_n)$  to denote elevation results under the inheritance and non-inheritance assumptions, respectively. The algorithm computes parsimonious representations for parental nodes according to the topology of the tree, proceeding from the leaves to the root in the manner which is similar to that described in [11] for a mathematical problem in bioinformatics.

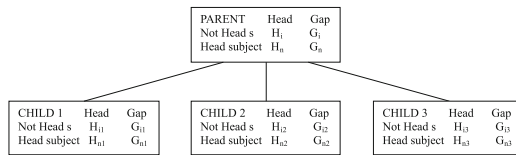


Fig. 4. Events in a parent-children system according to a parsimonious lift scenario

For the sake of simplicity, we present only a version of the algorithm for crisp clusters obtained by a defuzzification step. Given a crisp topic cluster  $S$ , and penalties  $h$ ,  $o$  and  $g$  for being a head subject, offshoot and gap, respectively, the algorithm is initialized as follows.

At each leaf  $l$  of the tree, either  $H_n = \{l\}$ , if  $l \in S$ , or  $G_i = \{l\}$ , otherwise. The other three sets are empty. The penalties associated are  $p_i = 0, p_n = o$  if  $H_n$  is not empty, that is, if  $l \in S$ , and  $p_i = g, p_n = 0$ , otherwise. This is obviously a parsimonious arrangement at the leaf level.

The recursive step applies to any node  $t$  whose children  $v \in V$  have been assigned with the two couples of  $H$  and  $G$  sets already (see Figure 4 at which  $V$  consists of three children):  $(H_i(v), L_i(v); H_n(v), L_n(v))$  along with associated penalties  $p_i(v)$  and  $p_n(v)$ .

(I) Deriving the pair  $H_i(t)$  and  $G_i(t)$ , under the inheritance assumption, the one of the following two cases is to be chosen depending on the cost:

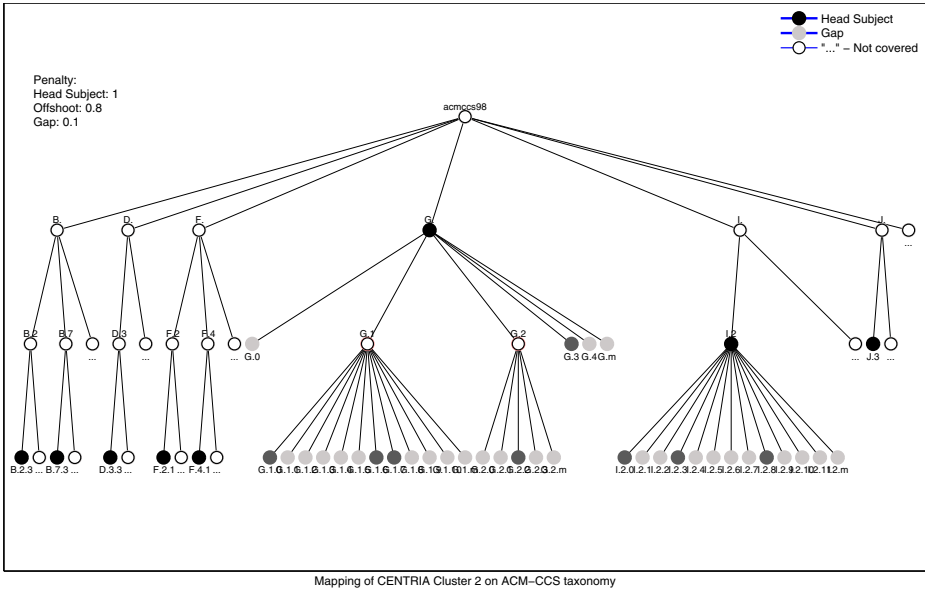
(a) The head subject has been lost at  $t$ , so that  $H_i(t) = \cup_{v \in V} H_n(v)$  and  $G_i(t) = \cup_{v \in V} G_n(v) \cup \{t\}$ . (Note different indexes,  $i$  and  $n$  in the latter expression.) The penalty in this case is  $p_i = \sum_{v \in V} p_n(v) + g$ ;

or

(b) The head subject has not been lost at  $t$ , so that  $H_i(t) = \emptyset$  (under the assumption that no gain can happen after a loss) and  $G_i = \cup_{v \in V} G_i(v)$  with  $p_i = \sum_{v \in V} p_i(v)$ .

The case that corresponds to the minimum of the two  $p_i$  values is returned then.

(II) Deriving the pair  $H_n(t)$  and  $G_n(t)$ , under the non-inheritance assumption, the one of the following two cases is to be chosen that minimizes the penalty  $p_n$ :



**Fig. 5.** Mapping of CENTRIA cluster 2 onto the ACM-CCS tree with penalties  $h = 1$ ,  $o = 0.8$  and  $g = 0.1$ : two head subjects along with 10 and 16 gaps, respectively

- (a) The head subject has been gained at  $t$ , so that  $H_n(t) = \cup_{v \in V} H_i(v) \cup \{t\}$  and  $G_n(t) = \cup_{v \in V} G_i(s)$  with  $p_n = \sum_{v \in V} p_i(v) + h$ ;
- or (b) The head subject has not been gained at  $t$ , so that  $H_n(t) = \cup_{v \in V} H_n(v)$  and  $G_n(t) = \cup_{v \in V} G_n(v)$  with  $p_n = \sum_{v \in V} p_n(v)$ .

After all tree nodes  $t$  have been assigned with the two pairs of sets, accept the  $H_n$ ,  $L_n$  and  $p_n$  at the root. This gives a full account of the events in the tree.

This algorithm leads indeed to an optimal representation; its extension to a fuzzy cluster is achieved through using the cluster memberships in computing the penalty values at tree nodes.

### 5 An Application to a Real World Case

Let us illustrate the approach by using the data from a survey conducted at the Centre of Artificial Intelligence, Faculty of Science & Technology, New University of Lisboa (CENTRIA-UNL). The survey involved 16 members of the academic staff of the Centre who covered 46 topics of the third layer of the ACM-CCS.

With the algorithm FADDI-S applied to the  $46 \times 46$  similarity matrix, two clusters have been sequentially extracted, after which the residual matrix has become definite negative (stopping condition (a)). Cluster 1 is of pattern recognition and its applications to physical sciences and engineering including images and languages, with offshoots to general aspects of information systems. In cluster 2, all major aspects of computational mathematics are covered, with an emphasis on reliability and testing, and with applications in the areas of life sciences. Overall these results are consistent with the informal

assessment of the research conducted in the research organization. Moreover, the sets of research topics chosen by individual members at the ESSA survey follow the cluster structure rather closely, falling mostly within one of the two.

Figure 5 shows the representation of CENTRIA's cluster 2 in the ACM-CCS taxonomy with penalties of  $h = 1$ ,  $o = 0.8$ , and  $g = 0.1$ .

## 6 Conclusion

We have proposed a novel method for knowledge generalization that employs a taxonomy tree. The method constructs fuzzy membership profiles of the entities constituting the structure under consideration in terms of the taxonomys leaves, and then it generalizes them in two steps. These steps are:

- (i) fuzzy clustering research topics according to their thematic similarities, ignoring the topology of the taxonomy, and
- (ii) elevating clusters mapped to the taxonomy to higher ranked categories in the tree.

These generalization steps thus cover both sides of the representation process: the empirical – related to the structure under consideration – and the conceptual – related to the taxonomy hierarchy.

Potentially, this approach could lead to a useful instrument for comprehensive visual representation of developments in any field of organized human activities.

## Acknowledgments

The authors are grateful to CENTRIA-UNL members that participated in the survey. Igor Guerreiro is acknowledged for developing software for the ESSA tool. Rui Felizardo is acknowledged for developing software for the lifting algorithm with interface shown in Figures 5. This work has been supported by grant PTDC/EIA/69988/2006 from the Portuguese Foundation for Science & Technology. The support of the individual research project 09-01-0071 “Analysis of relations between spectral and approximation clustering” to BM by the “Science Foundation” Programme of the State University – Higher School of Economics, Moscow RF, is also acknowledged.

## References

1. ACM Computing Classification System (1998), <http://www.acm.org/about/class/1998> (Cited September 9, 2008)
2. Bezdek, J., Keller, J., Krishnapuram, R., Pal, T.: Fuzzy Models and Algorithms for Pattern Recognition and Image Processing. Kluwer Academic Publishers, Dordrecht (1999)
3. Brouwer, R.: A method of relational fuzzy clustering based on producing feature vectors using FastMap. *Information Sciences* 179, 3561–3582 (2009)
4. Feather, M., Menzies, T., Connelly, J.: Matching software practitioner needs to researcher activities. In: Proc. of the 10th Asia-Pacific Software Engineering Conference (APSEC 2003), p. 6. IEEE, Los Alamitos (2003)

5. Gaevic, D., Hatala, M.: Ontology mappings to improve learning resource search. *British Journal of Educational Technology* 37(3), 375–389 (2006)
6. The Gene Ontology Consortium: Gene Ontology: tool for the unification of biology. *Nature Genetics* 25, 25–29 (2000)
7. Liu, J., Wang, W., Yang, J.: Gene ontology friendly biclustering of expression profiles. In: *Proc. of the IEEE Computational Systems Bioinformatics Conference*, pp. 436–447. IEEE, Los Alamitos (2004)
8. von Luxburg, U.: A tutorial on spectral clustering. *Statistics and Computing* 17, 395–416 (2007)
9. Miralaei, S., Ghorbani, A.: Category-based similarity algorithm for semantic similarity in multi-agent information sharing systems. In: *IEEE/WIC/ACM Int. Conf. on Intelligent Agent Technology*, pp. 242–245 (2005)
10. Mirkin, B.: Additive clustering and qualitative factor analysis methods for similarity matrices. *Journal of Classification* 4(1), 7–31 (1987)
11. Mirkin, B., Fenner, T., Galperin, M., Koonin, E.: Algorithms for computing parsimonious evolutionary scenarios for genome evolution, the last universal common ancestor and dominance of horizontal gene transfer in the evolution of prokaryotes. *BMC Evolutionary Biology* 3(2) (2003)
12. Mirkin, B., Nascimento, S.: *Analysis of Community Structure, Affinity Data and Research Activities using Additive Fuzzy Spectral Clustering*. Technical Report 6, School of Computer Science, Birkbeck University of London (2009)
13. Sato, M., Sato, Y., Jain, L.C.: *Fuzzy Clustering Models and Applications*. Physica-Verlag, Heidelberg (1997)
14. Skarman, A., Jiang, L., Hornshoj, H., Buitenhuis, B., Hedegaard, J., Conley, L., Sorensen, P.: Gene set analysis methods applied to chicken microarray expression data. *BMC Proceedings* 3(suppl. 4) (2009)
15. Shepard, R.N., Arabie, P.: Additive clustering: representation of similarities as combinations of overlapping properties. *Psychological Review* 86, 87–123 (1979)
16. Shi, J., Malik, J.: Normalized cuts and image segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 22(8), 888–905 (2000)
17. Thorne, C., Zhu, J., Uren, V.: *Extracting domain ontologies with CORDER*. Tech. Reportkmi-05-14. Open University, 1-15 (2005)
18. Yang, L., Ball, M., Bhavsar, V., Boley, H.: Weighted partonomy-taxonomy trees with local similarity measures for semantic buyer-seller match-making. *Journal of Business and Technology* 1(1), 42–52 (2005)
19. Zadeh, L.A.: Fuzzy sets. *Information and Control* 8, 338–353 (1965)
20. Zhang, S., Wang, R.-S., Zhang, X.-S.: Identification of overlapping community structure in complex networks using fuzzy c-means clustering. *Physica A* 374, 483–490 (2007)

# Anticipation as a Strategy: A Design Paradigm for Robotics

Mary-Anne Williams, Peter Gärdenfors,  
Benjamin Johnston, and Glenn Wightwick

Innovation and Enterprise Research Laboratory  
QCIS, University of Technology, Sydney  
Mary-Anne@TheMagicLab.org

**Abstract.** Anticipation plays a crucial role during any action, particularly in agents operating in open, complex and dynamic environments. In this paper we consider the role of anticipation as a strategy from a design perspective. Anticipation is a crucial skill in sporting games like soccer, tennis and cricket. We explore the role of anticipation in robot soccer matches in the context of reaching the RoboCup vision to develop a robot soccer team capable of defeating the FIFA World Champions in 2050. Anticipation in soccer can be planned or emergent but whether planned or emergent, anticipation can be designed. Two key obstacles stand in the way of developing more anticipatory robot systems; an impoverished understanding of the “anticipation” process/capability and a lack of know-how in the design of anticipatory systems. Several teams at RoboCup have developed remarkable preemptive behaviors. The CMU Dive and UTS Dodge are two compelling examples. In this paper we take steps towards designing robots that can adopt anticipatory behaviors by proposing an innovative model of anticipation as a strategy that specifies the key characteristics of anticipation behaviors to be developed. The model can drive the design of autonomous systems by providing a means to explore and to represent anticipation requirements. Our approach is to analyze anticipation as a strategy and then to use the insights obtained to design a reference model that can be used to specify a set of anticipatory requirements for guiding an autonomous robot soccer system.

**Keywords:** Design, Strategy, Decision, Anticipation, Perception, Behavior.

## 1 Introduction

This paper seeks to improve autonomous robot system design by focusing on the development of a mobile robot soccer system as a benchmark problem. Anticipation is a crucial strategy for robots that must reason and act in real time in open, complex and dynamic environments, and anticipation in a competitive team sport like robot soccer has all the characteristics of anticipation that need to be designed. Anticipation involves reasoning or acting preemptively on the basis of past and present experiences, e.g. inward and outward facing perceptions. Robot behaviors involve the enactment of information processing, and based on mathematical equations and simple data

structures like lookup tables; the execution of elaborate predictive simulations; and morphological properties and responses.

The vision of RoboCup is to develop a team of robots that can defeat the FIFA World Champions in 2050. Seeing a robot anticipate during a soccer game today is exciting. It will be difficult, perhaps impossible, for a team of robots without the ability to anticipate on the fly to beat a resourceful, imaginative and cunning human soccer team. It is expected that robot soccer players of 2050 will be physically stronger, faster and more agile than human players with superhuman capabilities in sensing, information sharing, and cooperation and collaboration. However, without the ability to anticipate efficiently and effectively the robots will be at a significant disadvantage even against a physically inferior team of biological humans.

A team that can anticipate will outclass any team that cannot, simply because there are physical limits to an agent's response time and as a result purely reactive behaviors will always lag the play, instead of being abreast or ahead of it. *Anticipating a play* involves interpreting/reading and responding to it in a timely fashion; there is no point attempting to trap or kick a ball that has already passed you. The ability to anticipate allows teams to position themselves advantageously in a game. If players can anticipate, they can pass, intercept, and adopt defensive attacking strategic moves. Passing a ball is an essential capability in soccer and a team that can pass, as well as anticipate the opposition's passes will outclass a team that can only pass but not anticipate opposition passes - unless, of course, it can maintain possession, which is not realistic in any game. Given the enabling role anticipation plays, a key design issue in robot soccer is how to develop autonomous robotic systems that can anticipate efficiently and effectively. Human soccer players effortlessly anticipate the motion of the ball and future position of players, but without robust anticipatory capabilities, it is hard to imagine a robot team out-playing the top human team in 2050.

Section 2 explores anticipation as a strategy drawing on insights from the extensive study and analysis of strategy in management and organizational behavior. Section 3 identifies the key role anticipation plays in competitive sports. Section 4 highlights the need for using anticipatory behaviors in robot soccer. Section 5 provides a means to identify, specify and represent the requirements of anticipative behavior from the perspective of using it as a strategy, which serves as an important first step in the design of autonomous robot soccer team that can anticipate.

## 2 Anticipation as a Strategy

Anticipation is a complex concept, and it is often confused with prediction. Anticipation and prediction are different, albeit related, processes and capabilities. Anticipation is similar to prediction in that it can involve forecasting something in the a future; world state, a perception, a need, a want. On the other hand, anticipation from a design perspective is quite different to prediction in four crucially important ways.

First, anticipation demands a response. Prediction can be used to forecast the path of the ball after it has been kicked while anticipation involves a response of some kind such as a decision to seize possession of the ball when it gets close enough. When an agent anticipates it takes steps to prepare for an (internal or external) event before it happens. It is important to understand that anticipation is not about predicting



outcomes or unknowns but determining what the agent should do in preparation for something that may happen (in the future) or the new discovery that something has already happened (in the past). It would be reasonable to say that anticipation can rely on or use prediction, but not conversely.

Second, prediction in ordinary parlance presupposes an underlying causal model, while, in contrast, anticipation need not. Anticipation might be based on something much more basic than a predictive model something like a wild guess, a random starting place in a search strategy, a simple cue, a recognizable pattern, a logical abduction, rough intuition, or an explanatory model i.e. a model with explanatory power, but lacking predictive power. The line between prediction and anticipation is can be murky because one could argue that predictive models routinely use random estimates, however, a random estimate alone for most purposes would not rate as a reliable predictive model it would be just playing the odds, so to speak. Moreover, anticipation involves a response so the response to the random estimate is what counts and the subsequent performance of the response.

Third, prediction is often objective and its performance can be measured objectively. A valuable and measurable feature of a predictive model is that it reliably predicts outcomes in such a way that the results (e.g. predictions) are (scientifically) repeatable to a measurable/describable degree of approximation. Anticipation, however, tends to allow more scope for subjectivity. Anticipation typically takes aspects of a given agent into consideration because agents anticipate while a mechanical devices without any form of agency predict, e.g. a wristwatch. In contrast, anticipating the weather involves making the decision of whether to take an umbrella, and therefore depends on what activities are planned, how they are intended to be performed and possibly on state of being, such as physical health. A foreign exchange investment strategy does not solely depend on predicting the relative value of currencies, but on the relevant business' situation, resources and objectives. This subjective property becomes central in robot design because it provides the useful insight that designing for anticipation should be done relative to agent experiences; it highlights the need to consider robots as independent agents interacting in the world and anticipating on the basis of its own experience.

Fourth, anticipating is often easier than prediction because time horizons can be fleeting, particularly when there is scope for feedback and interactivity. Developing a predictive model typically requires deep causal understanding developed over extensive experience or using sophisticated knowledge. Contrast the difficulty of solving Euler's three-body problem in physics with an amateur juggler who can effortlessly juggle three or more balls without difficulty. Furthermore, there are many problems and scenarios for which we just do not have precise predictive models, e.g. long range weather, yet we are able to anticipate the weather for the purpose of making various plans, indeed we often make contingent plans that cover the range of expected possibilities using our past experience. For example, when organizing a picnic in several months time one would not study solar flare activity or current trends in the solar wind, as one would in weather prediction weather prediction; but instead anticipate the general seasonal weather trends and plan for contingencies.

Anticipation is an important trait for many systems including biological (e.g. plants and animals), organizational (e.g. businesses and universities), and designed agents (e.g. software systems and robots). Developing anticipatory capabilities and management

designs for organizations is similar in many important respects to that for autonomous robots. Both are distributed systems, and as a result organizational leaders and robot designers face similar challenges including how to develop capabilities for anticipation in friendly, competitive and adverse environments. Designs for organizations and robots share many important characteristics including the need to align actions and behaviors across the whole distributed system (organization or robot). Poor performing anticipation capabilities become a liability for a system, while efficient and effective anticipation can lead to success. We use Mintzberg's 5P approach [9] to strategy as a tool for exploring, analyzing and describing key characteristics of *anticipation as a strategy* as follows:

1. *Anticipation as a plan*; purposeful, intentional but also agile, adaptable and flexible. Plans are typically a series of actions that are sometimes conceptualized as a pathway from one state (e.g. current state) to another (e.g. desired future state) in the pursuit of goals. Planning in this sense has been well studied in AI.
2. *Anticipation as a ploy*; a specific "maneuver" designed to outwit an opponent or competitor in adversarial or competitive contexts. Dodging and diving in soccer are examples of ploys.
3. *Anticipation as a pattern* of actions and behavior; anticipation may emerge from behaviors. We will use anticipation patterns to describe emergent anticipation, e.g. the emergent opportunistic pass versus the deliberative, carefully crafted and synchronized pass.
4. *Anticipation as a subjective position*; position is relative to a competitive context like a position of advantage in a game. A robot in possession of the ball is in a position of advantage; it can dictate the play and exercise indirect control over other players and various outcomes to some extent.
5. *Anticipation as a perspective*; anticipatory behavior is often undertaken with respect to a certain frame. One player or several players might be involved. Other perspectives include egocentric (vision stream), allocentric views (world model).

Strategy is distinct from tactics. We use tactics as components of anticipation strategies herein and consider a tactic to be single, possibly non-decomposable, behaviour deployed for a reason, and potentially different reasons in different strategies. A strategy can dynamically link tactics together for a larger purpose creating dynamic capabilities. A dynamic capability is the ability to integrate, build, and reconfigure internal and external competences to address rapidly changing environments [41, 41]. The dynamic characteristic of a strategy allows for both deliberately planned strategic moves and emergent strategic moves to be composed on the fly.

Strategy can have different meanings in different contexts. For example, in game theory, a *strategy* refers to one of the options from which a player can choose where every player has a set of possible strategies. A strategy specifies the actions that will occur in each contingent state of the game e.g. if the opponent undertakes action  $a_1$ , then take action  $a_2$ ; if the opponent undertakes  $a_3$ , then action  $a_4$ . Strategies in game theory scenarios range from random to deterministic.

There is considerable work in cognitive science on anticipation. Pezzulo [12] provides empirical evidence indicating that anticipatory representations are involved in several low and high-level cognitive functions such as attention, motor control, planning and goal-oriented behaviour. Sjolander [39] points out that the ability to

anticipate future states is a major evolutionary and cognitive advance. He considers an agent able to anticipate if it is able to take the effects of its own actions into account at different future timescales. Rosen [19] argues that anticipation is chemically and physically built-in to life itself, i.e. all living creatures exhibit anticipatory behavior.

### 3 Anticipation in Sport and Games

In tennis, the return of a serve is successful only through anticipatory mechanisms on the part of the defensive player [11]. Anticipation is a strategy used by athletes “to reduce the time they take to respond to a stimulus”, e.g. to get into a position to reach the ball for a return of serve. There are many situations in sports and games where even an instantaneous reaction would not provide enough time for a player to get to where he could be had he anticipated a given play. A player’s anticipatory capabilities need only be better than random before they start to pay dividends (indeed in some cases random behaviours more advantageous than no action at all), and obviously the higher the success rate the more prudent and beneficial anticipation will be as a strategy. Scope for improvement is important; practice can have a dramatic affect on players’ ability to anticipate.

In order to anticipate a player must pay attention and interpret patterns of activity/behavior that can act as (perceptual) cues and signals, e.g. looking in a certain direction before kicking a ball can indicate where the ball is headed (or not). Perceiving attention in others can be used to anticipate their intentions and future actions.

Anticipation is apparent in some existing RoboCup soccer team systems, e.g. diving as the ball approaches [CMU AIBO team] and dodging opposition robots [UTS AIBO team]. However, these examples are carefully crafted add-ons rather than a natural evolution of an anticipation enabling architecture and design. Nonetheless they provide powerful and compelling evidence that anticipation is something to aim for in mobile robot design and importantly that it can be achieved. These proof-of-concept exemplars, also highlight one of the main design challenges, namely how to identify and specify anticipatory behavior.

We advocate a new design paradigm where robot soccer systems are designed with anticipation as the underlying driving strategy from scratch where all aspects of the robot system are grounded in anticipation capabilities. In soccer it is well known that the team that is able to anticipate the opposition’s moves will be significantly advantaged because they will have the opportunity to respond in a timely fashion and also be able to develop counter strategies, rather than simply being at the mercy of the opposition. Somewhere down the track RoboCup teams will need to develop robust techniques for designing and deploying anticipatory and adaptive game play; game plays that respond to change and the game dynamics. At present few, if any, teams adapt their strategy on the fly during a match in response to opposition behaviors, except through human intervention.

Deceptive play is prevalent in games and it is difficult to anticipate; deliberate deceptive play by an opposition team requires sophisticated interpretation and responses. For example, a player might pretend not to notice a deceptive play for a while in order to first learn more about it and second to lull the opposition into the

false belief that the deception is working and then surprise them at a crucial time when they least expect it. In terms of robot development it is important to note that humans are the only species that appear to routinely use deception. It is well known that an agent requires a so-called *theory of mind* in order to deceive, and presumably an ability to deceive is needed to recognize deception in others. That does not mean that robots without a theory of mind cannot deal with deception (ignorance can be bliss), but it probably makes effective anticipatory behaviors more difficult to develop. On the other hand, unlike humans robots can share their raw sensory and perceptual data directly, so building robots with a primitive theory of mind is clearly possible; it is more a question of the level of sophistication and capability that can be achieved.

Good players learn to notice, pay attention and assess certain cues early in a play, consider possible outcomes and determine more effective responses than if they waited to observe what would happen. Anticipatory systems look for indications of what may eventuate so that they can prepare for it, if possible. Anticipation and attention are intimately interlinked.

Anticipation allows players to position themselves earlier than if they simply waited until they sensed the impending changes directly, e.g. imagining how things might change and responding ahead of time may give more time to enact a higher quality response. Obviously, there are inherent dangers in anticipating, if the player gets it wrong e.g. a goalie goes left instead of right when a player kicks the ball at the goal. But surprisingly the relative cost can be small because the cost is not the difference between getting it right and getting it wrong, but not acting at all and undertaking the wrong action. For example, standing flat-footed and waiting until the ball is upon you and then reacting is typically not better than being in motion and going the wrong way. Furthermore, engaging in anticipation provides extraordinary learning opportunities, which can be seen as worth the cost especially in light of the fact that making mistakes and recovering from them can improve learning performance. When it comes to designing for anticipation it is important to realize the underlying cost and risk relationship at work in a particular scenario and context. In general the upfront costs and opportunity costs for not anticipating (i.e. doing nothing) is greater than anticipating incorrectly sometimes, and importantly the rewards for sometimes getting it right can be substantial, e.g. some of the best strikers in the history of soccer average less than a 35% success rate from goal attempts; Cristiano Ronaldo of Manchester United and now Real Madrid fame has a professional career average of less than 20%; he is one of the highest goal scorers per game but his percentage of success is surprising low.

People improve their anticipatory skills with practice; they improve their ability to imagine possibilities and to improve their response times to cues through drills, practice and experience playing real games. The objective of drills and practice routines is to improve performance, e.g. improve reaction time to stimuli. Soccer drills include ball control, kicking, passing and deceptive plays.

## 4 Anticipation in Robots

Anticipation plays a crucial role in most physical activities especially competitive sport. Anticipation can be deliberately planned, e.g. the execution of a pre-planned

synchronized pass; anticipation can also emerge, e.g. an opportunistic pass based on robot behavior that enacts kick to teammate; it can also occur by accident, e.g. a pass that just happens as a game unfolds as a result of behaviors like “kick the ball into open spaces” and “if I do not have possession of the ball move into an open space”. Whether planned, emergent or accidental, anticipation can be designed. Anticipation is an innate ability in humans, and anticipation-enabled robots can achieve more than robots that wait for events to occur. A soccer player that cannot anticipate will be caught flat-footed and respond too late to relevant stimuli to be competitive.

An important family of anticipatory behaviors is decision-making using predictions, expectations, or beliefs about the future. Robot behaviors like kicking towards the goal and returning to defensive positions and bearing provide compelling examples of how efficient approximate predictive models like Kalman filters can be used to support anticipation, leading to significant advantage in a robot soccer game. Logic based specifications are also useful in developing anticipatory behaviors. For example, “if a teammate is in the clear, kick the ball so that it lands in his expected path”. New work on commonsense reasoning that unites logic and simulation [6] can be used to model the rule above by evaluating expected paths using simulation rather than having to develop a reversible rational model. Case-based reasoning can also be used effectively to respond to internal and external cues and their combinations [22, 23] – determining how to respond is normally informed by where a robot is and what he is doing (stationary, moving in a certain direction) and what is happening around him (where the ball is going, or likely to be by the time he reaches it given his current trajectory and velocity).

Anticipatory behavior can be reactive or made up of reactive elements. Reactive systems have the property that there is a single/fixed next action for each current state. Reactive decision-making is based on the current state of the environment and the current state can hold cues and clues for future states or other representations of relevant aspects of the future. A robot employing anticipation might predict a future state of the environment (e.g. the ball will reach me soon) as a means to make a decision (e.g. to choose an action). Reactive decision-making is a response to existing conditions. Proactive decision-making, on the other hand, takes possible future events/states/goals into account, implicitly or explicitly.

Robert Rosen [19] defined an anticipatory system as a system that contains a predictive model of itself and/or its environment, which allows it to change state at an instant in accord with the model's predictions pertaining to a later instant. This definition makes some strong assumptions. For us anticipation can be less orchestrated. It does not necessarily require a predictive model, nor does it need to pertain to a later instant, i.e. “the future”, even if it takes place over time.

Realistically robots need to be able to improve performance with practice. This essentially means they must be able to interpret their own experience and find ways to improve their skills possibly in collaboration with a coach. Reaction time itself is an inherent ability, but overall response time can be improved by practice. Machine learning in robot locomotion [8, 23] provides a good example of how useful practice is in improving performance. Coaches, athletes and robot soccer player designers need to analyze the type of skill to practice and determine where overall response gains can be made, e.g. detecting relevant cues, developing set plays, focusing attention, switching attention.

Machine learning techniques are commonly used in robot soccer, e.g. robots can time themselves as they walk from one side of the field to the other as they improve the performance of certain walks such as speed and stability using reinforcement learning [8] and genetic algorithms [23]. Interestingly robots suffer from wear and tear just like humans; robots are typically at a distinct disadvantage because, currently, they do not have self-healing mechanisms that biological creatures enjoy.

## 5 Designing Robot Anticipatory Behaviors

Anticipation as a strategy can be a powerful driving force in designing and developing mobile robotic agents. The importance of anticipation and robot design is not a new idea. However, practical approaches to anticipatory design are lacking. Our proposal draws inspiration from the realization that managing an organization is similar to managing an autonomous robot since they are both distributed systems that make and enact decisions in behaviours, and as a result we were able to use strategy concepts from the field of strategic management to inform robot design.

There are two main approaches to developing strategies; prescriptive and descriptive. The prescriptive approaches seek to identify the behaviors an agent should adopt based on its assessment of its current situation. Descriptive approaches, on the other hand, focus on the reasons an agent is where it is to determine what it should do next. We identify several relevant prescriptive and descriptive approaches that can be used in a design methodology for robots as follows:

- Design approach uses a best possible fit approach based on matching strategies.
- Planning approach is a formal stepwise process from an analysis of the situation to the development and exploration of alternative scenarios.
- Positioning approach is based on an analytical process that places the agent within the specific context and using the *in situ* insights determines how the organization can improve its competitive positioning within that context.
- Entrepreneurial approach is less structured, and explores the imagination of the agent and its designers.
- Cognitive approach considers strategy formation as a mental process, and analyzes how agents perceive patterns and process information.
- Learning approach regards strategy formation as an emergent process, where the management of an organization pays close attention to what works and doesn't work over time, and incorporates these 'lessons learned' into their overall plan of action.
- Power approach to strategy development is a process of negotiation between power holders; in a robot design the control system can be thought of as a collection of power holders.
- Environmental approach is based on a reactive process understanding: a response to the challenges imposed by the external environment; in this approach it is assumed that the environment dictates to the agent and the agent has limited forms of choices to environmental stimuli.

**Table 1.** Anticipation as Strategy Design Specification Representation of a preplanned *Dodge* maneuver and an emergent *Pass*

<b>Anticipation Frame Name: Dodge</b>
<b>Type:</b> Deliberate and completely specified.
<b>Kind:</b> Tactic; single behavior package
<b>Plan:</b> Used when robot is in possession of the ball and obstructed
<b>Ploy:</b> Attacking Ploy: Used to line up a goal better Defensive Ploy: Used to intercept a ball in play or about to be kicked
<b>Pattern:</b> Anticipatory passes could emerge from this behavior. Passing robots dodge to pass, receiving robot dodges to catch.
<b>Position:</b> Used to gain better field position, used to gain clearer sight of goal, used to find teammates in the clear, used to steal ball possession
<b>Perspective:</b> Offensive move when in possession of the ball. Defensive if opposition has the ball.
<b>Decision:</b> Offensive: Can it evoke a dodge before kicking? Defensive: Can it evoke a dodge after finding the ball?
<b>Cue:</b> - Possession of the ball and facing goal; ball moving towards robot
<b>Attention:</b> In state of possession and obstacles immediately in front; focus on ball
<b>Response:</b> $n$ steps to left or right, while maintaining possession
<b>Time:</b> No explicit representation of time
<b>Performance:</b> Percentage of successful kicks after the dodge
<b>Improvement:</b> Distinguish stationary from moving object, decision to move left or right
<b>Risk:</b> Fumble and loss of possession, interception
<b>Reward:</b> Improved field position and increase chance of scoring
<b>Traps:</b> No detectable traps
<b>Prerequisite Capabilities:</b> Basic infrastructure: locomotion, vision, behavior control

<b>Anticipation Frame Name: Pass</b>
<b>Type:</b> Emergent Only the kicking robot is aware of the pass. It aims to place the ball in front of another teammate who will scoop it up and kick. Specified for one robot only. No explicit communication between robots. Relies on kicking robots attention to other robot's place and bearing, and the receiving robots basic instinct of <u>grabbing the ball when it is close.</u>
<b>Kind:</b> Tactic; single behavior package
<b>Plan:</b> Kick in front of players path
<b>Ploy:</b> Used for direct attacking strike; used for field advantage;
<b>Pattern:</b> Clearance kicks, backwards kicks
<b>Position:</b> Used to gain better field position, used to gain advantage over a faster opposition team.
<b>Perspective:</b> Offensive move to push ball forward and maintain possession; Defensive when used to scramble the opposition.
<b>Decision:</b> Is there a location of target for kick?
<b>Cue:</b> - Clear line of sight to teammate moving or facing forward, or looking at ball facing kicking robot.
<b>Attention:</b> Ball possession and control, and view of target player.
<b>Response:</b> Kick to target location
<b>Time:</b> No explicit representation of time
<b>Performance:</b>
<b>Improvement:</b> Target location
<b>Risk:</b> Fumble and loss of possession, interception
<b>Reward:</b> Improved field position and increase chance of scoring
<b>Traps:</b> No detectable traps
<b>Prerequisite Capabilities:</b> Basic infrastructure; locomotion, vision, behavior control

We take a hybrid cognitive-entrepreneurial learning design approach and focus on identifying and representing anticipation capabilities and behaviors. Practical robotic system development involves mapping a set of requirement specifications to a robotic platform, architecture and design. Therefore we develop an *Anticipation as Strategy Design Specification Representation* (ASDSR) framework using the frames illustrated in Table 1, below. ASDSR frames provide a way to describe anticipation capabilities and behaviors. A design is as good as the quality of its requirements specification; we focus on requirements analysis and representation. ASDSR frames encapsulate the proposed understanding of anticipation presented in the previous sections, and are used to drive requirements analysis and to describe the anticipatory behaviors to be designed. It describes anticipatory strategies by representing and describing characteristics, capabilities and affordances.

Machine learning methods can be used to build anticipatory capabilities. Robots have been shown to be able to learn to anticipate future rewards and punishments caused by current actions [25]. Importantly, Balkenius and Hulth [1, 44] showed anticipation actually improved learning performance where attention can be guided proactively to collect relevant information in order to act effectively.

The ASDSR is used to identify relevant “things” that a robot should pay attention to in each anticipation strategy. A robot has limited computational resources and just like people, cannot pay attention to everything – there are an infinite number of potential “things” to which attention could be directed. It turns out that if robot reaction time is measured as the interval between the presentation of a stimulus and the initiation a hardware response to that stimulus, then a primary factor affecting a response is the number of possible stimuli requiring a response that are presented. Hick [7] discovered that reaction time in people increases proportionally to the number of possible alternatives (or to the amount of information that must be processed in order to respond), until a point at which the response time remains constant despite the increases in possible responses. Hicks found that reaction time =  $a + b (\log_2 N)$ , where  $a$  and  $b$  are constants and  $N$  is the number of stimulus-response alternatives [7].

An important anticipation-as-strategy design principle says that response times will improve if the number of stimuli attended is less; the optimum being a single simple response. Practice can be seen to be a way to reduce the number and complexity of stimuli a robot learns to pay attention to things that really matter.

In addition to these insights in studies into attention and reaction time Schmidhuber [22] showed how modifying the error back propagation algorithm can be used to change neural network weights so that to the mismatch between anticipated states and states actually experienced in the future decrease over time. Schmidhuber was also able to define a notion of *curiosity* for agents as a measure of the mismatch between expectations and future experienced reality. As a result agents can be designed to monitor and control their own anticipation driven curiosity. This feature would allow robots to explore their own physical capabilities as well as their teammates, and to seek new soccer playing experiences.

## 6 Discussion

The ability to anticipate in soccer is of crucial importance for intelligent agents like people and autonomous robots. It is simply not possible to play soccer well against a



competent team without anticipating the behavior of players and the ball. In this paper we explored the need to develop anticipation as a strategy for robot design using robot soccer as a benchmark problem. Anticipating and preparing for an eventuality is different to predicting the eventuality, although clearly prediction and anticipation can be related. Anticipation involves imagination and curiosity. Anticipating that a player will kick the ball forward is different from predicting the ball's path. It requires an awareness of the player's intentions, capabilities and skills. Understanding the physics of the ball is not necessary just the understanding that if one wants to be close to the ball one should move down the field in the direction it was kicked. An accurate prediction of the balls behavior is not necessary because the robot can perceive the ball as it rolls along, particularly if it possesses object permanence and can anticipate that if the ball goes behind another robot it has not vanished or no longer exists. Moreover, he can use information about his own body e.g. how quickly he has to move his head to keep the ball in view as a means to judge how fast the ball is moving.

Anticipation is much more open as a capability and process than prediction, it is about awareness of the possibilities and the means to interpret what is going on within and around you by focusing attention on the minimal number of things that matter or that could have an impact on you or your objectives. RoboCup robot soccer teams can start with simple anticipation models and over time increase their sophistication as the community develops a new understanding of how to design and build anticipatory capabilities in autonomous soccer playing robots. We provided a new design framework and frame representation that can help guide the development process by helping a designer gather and specify the relevant requirements that are needed when building-in anticipation as a strategy for design.

## References

1. Balkenius, C.: *Natural Intelligence in Artificial Creatures*, p. 37. Lund University Cognitive Studies (1995)
2. Card, S.K., Moran, T.P., Newell, A.: *The Psychology of Human-Computer Interaction* (1983)
3. Cockburn, A., Gutwin, C., Greenberg, S.: A predictive model of menu performance. In: *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (2007)
4. Davis, B.: *Physical Education and the Study of Sport*, ISBN 0 7234 31752
5. Dintiman, G.: *Sports Speed*, ISBN 0 88011 607 2
6. Johnston, B., Williams, M.: *Comirit: Commonsense Reasoning by Integrating Simulation and Logic*. In: *Artificial General Intelligence 2008*, pp. 200–211 (2008)
7. Kent, M.: *Sports Science & Medicine*. Oxford Dictionary of Sports Science & Medicine (2007)
8. Kim, M.S., Uther, W.: Automatic gait optimization for quadruped robots. In: *Proceedings of the 2003 Australasian Conference on Robotics and Automation* (2003)
9. Mintzberg, H., Ahlstrand, B., Lampel, J., Safari, S.: *A Guided Tour Through the Wilds of Strategic Management*. The Free Press, New York (1998)
10. McArdle, W.D., et al.: *Essentials of Exercise Physiology*, ISBN 0 683 30507 7
11. Nadin, M.: *Anticipation-The End Is Where We Start From*. Lars Müller Publisher (2003)
12. Pezzulo: Anticipation and Future-Oriented Capabilities in Natural and Artificial Cognition, pp. 257–270. Springer, Heidelberg (2007)

13. Seow, S.C.: Information Theoretic Models of HCI: A Comparison of the Hick-Hyman Law and Fitts' Law. *Human-Computer Interaction* 20(3), 315–352 (2005)
14. Welford, A.T.: *Fundamentals of Skill*. Methuen, 61–65 (1968)
15. Pezzulo, G.: Anticipation and Future-Oriented Capabilities in Natural and Artificial Cognition. In: Lungarella, M., Iida, F., Bongard, J.C., Pfeifer, R. (eds.) *50 Years of Artificial Intelligence*. LNCS (LNAI), vol. 4850, pp. 257–270. Springer, Heidelberg (2007)
16. Pezzulo, G., Butz, M.V., Castelfranchi, C., Falcone, R. (eds.): *The Challenge of Anticipation: A Unifying Framework for the Analysis and Design of Artificial Cognitive Systems*. LNCS (LNAI), vol. 5225. Springer, Heidelberg (2008)
17. MacLeod, A., Conway, C.: Well-being and the anticipation of future positive experiences: The role of income, social networks, and planning ability
18. Kunde, W.: No anticipation–no action: the role of anticipation in action and perception. *Cognitive Processing* 8(2) (June 2007)
19. Rosen, R.: *Anticipatory Systems*. Pergamon Press, Oxford (1985)
20. Sutton, R.S., Barto, A.G.: *Reinforcement Learning: An Introduction*. MIT Press, Cambridge (1998), A Bradford Book
21. Speed, T.: McNab, Advanced Studies in Physical Education and Sport. In: Beashel, P., et al
22. Schmidhuber: Curious model-building control systems. In: *Proc. International Joint Conference on Neural Networks*, Singapore, vol. 2, pp. 1458–1463. IEEE, Los Alamitos (1991)
23. Anshar, M., Williams, M.: Extended Evolutionary Fast Learn-to-Walk Approach for Four-Legged Robots. *Journal of Bionic Engineering* 4(4), 255–264 (2007)
24. Karol, A., Nebel, B., Stanton, C., Williams, M.-A.: Case-based Game-Play in the RoboCup Four-Legged League. In: Polani, D., Browning, B., Bonarini, A., Yoshida, K. (eds.) *RoboCup 2003*. LNCS (LNAI), vol. 3020, pp. 739–747. Springer, Heidelberg (2004)
25. *MindRACES: From Reactive to Anticipatory Cognitive Embodied Systems* (2004), <http://www.mindraces.org>
26. Ros, Arcos, J.L., Lopez de Mantaras, R., Veloso, M.: A case-based approach for coordinated action selection in robot soccer. In: *Artificial Intelligence* (2009)
27. Butz, M.V., Sigaud, O., Gérard, P.: Anticipatory Behavior: Exploiting Knowledge about the Future to Improve Current Behaviour. In: Butz, M.V., Sigaud, O., Gérard, P. (eds.) *Anticipatory Behavior in Adaptive Learning Systems*. LNCS (LNAI), vol. 2684, pp. 1–10. Springer, Heidelberg (2003)
28. Camacho, E., Bordous, C.: *Model Predictive Control*. Springer, Berlin (1998)
29. Hoffmann, J.: Anticipated Behavioral Control. In: Butz, M.V., Sigaud, O., Gérard, P. (eds.) *Anticipatory Behavior in Adaptive Learning Systems*. LNCS (LNAI), vol. 2684, pp. 44–65. Springer, Heidelberg (2003)
30. Luhmann, N.: *Social Systems*. Stanford University Press, Stanford (1995)
31. Mulcahy, N., Call, J.: Apes Save Tools for Future Use. *Science* 312, 1038–1040 (2006)
32. Nadin, M.: *Anticipation. The End is Where We Start From*. Lars Mueller Publishers (2004)
33. Negenborn, R.R., Schutter, B.S., Hellendova, J.: *Multi Agent Model Predictive Control. A Survey*. Deft: Deft Center for Systems and Control, Deft University of Technology, Technical Report 04-010 (2004)
34. Poli, R.: The Ontology of What is Not There. In: Malinowski, J., Pietruszczak, A. (eds.) *Essays in Logic and Ontology. Essays dedicated to Jerzy Perzanowski*, Rodopi, pp. 73–80 (2006)
35. Raby, C.R., Alexis, D.M., Dickinson, A., Clayton, N.S.: Planning for the Future by Western Scrub-jays. *Nature* 445, 919–921 (2007)

36. Riegler, A.: Whose Anticipations? In: Butz, M.V., Sigaud, O., Gérard, P. (eds.) *Anticipatory Behavior in Adaptive Learning Systems*. LNCS (LNAI), vol. 2684, pp. 11–22. Springer, Heidelberg (2003)
37. Rosen, R.: *A Relational Theory of Biological Systems*. *Bulletin of Mathematical Biophysics*, 245–260 (1958)
38. Rosen, R.: *Anticipatory Systems. Philosophical, Mathematical and Methodological Foundations*. Pergamon Press, Oxford (1985)
39. Rosen, R.: *Essays on Life Itself*. Columbia University Press, New York (2000)
40. Rosen, R.: *Fundamentals of Measurement and Representation of Natural Systems*. North Holland, New York (1978)
41. Sjölander, S.: Some cognitive breakthroughs in the evolution of cognition and consciousness & their impact on the biology of language. *Evolution and Cognition*, 3–11 (1995)
42. Tolman, E.C.: *Purposive Behavior in Animals and Men*. Appleton, New York (1932); Tolman, E.C.: There is More Than One Kind of Learning. *Psych. Review*, 144–155 (1949)
43. Teece, D., Pisano, G., Shuen, A.: Dynamic Capabilities and Strategic Management. *Strategic Management Journal* 18(7), 509–533 (1997)
44. Eisenhardt, K., Martin, J.: Dynamic Capabilities: What are they? *Strategic Management Journal* (21), 1105–1122 (2000)
45. Pfeifer, R., Bongard, J.C.: *How the Body Shapes the Way We Think: A New View of Intelligence*. MIT Press, Cambridge (2006)
46. Balkenius, C., Hulth, N.: Attention as Selection-for-Action: A Scheme for Active Perception. In: *Proceedings of EUROBOT 1999*, ETH, Zürich (1999)
47. UTS Dodge,  
[http://unleashed.it.uts.edu.au/assets/  
UTSUnleashedDodgeRoboCup2004.mov](http://unleashed.it.uts.edu.au/assets/UTSUnleashedDodgeRoboCup2004.mov)

# Modular Logic Programming for Web Data, Inheritance and Agents

Isambo Karali

Department of Informatics and Telecommunications  
University of Athens  
Panepistimiopolis  
Ilissia, Athens GR-15784  
izambo@di.uoa.gr

**Abstract.** The Semantic Web provides a framework and a set of technologies enabling an effective machine processable information. However, most of the problems that are addressed in the Semantic Web were tackled by the artificial intelligence community, in the past. Within this period, Logic Programming emerged as a complete framework ranging from a sound formal theory, based on Horn clauses, to a declarative description language and an operational behavior that can be executed. Logic programming and its extensions have been already used in various approaches in the Semantic Web or the traditional Web context. In this work, we investigate the use of Modular Logic Programming, i.e. Logic Programming extended with modules, to address issues of the Semantic Web ranging from the ontology layer to reasoning and agents. These techniques provide a uniform framework ranging from the data layer to the higher layers of logic, avoiding the problem of incompatibilities of technologies related with different Semantic Web layers. What is more is that it can operate directly on top of existent World Wide Web sources.

World Wide Web, Semantic Web, Logic Programming, Modularity.

## 1 Introduction

Nowadays, the Web comprises a large amount of information sources with varying quality, syntactical diversity but often semantical relevance. The Semantic Web [36] framework and standards emerged from an effort to improve the utilization of the web potential. To achieve this, the Semantic Web provides a set of technologies for data structuring, description and sharing as well as provision or means for reasoning, knowledge support and trust on sources. The approach is structured in layers [37] each addressing a higher problem. XML [43], RDF [29], RDF Schema [30] and OWL [28] are the languages used to express the requirements on the data and metadata. Various initiatives, such as RuleML [32], have been established to provide a means for reasoning and higher level capabilities.

Logic programming [24] and its extensions have been already used in various approaches in the Semantic Web or the traditional Web context. For instance,

in [25], web pages were treated as logic programs. Currently, a lot of discussions concern the issue of “negation as failure” encountered in logic programming when applied to the Semantic Web context, e.g [14]. For quite many Web application areas, though, e.g. [35,19], we need to store only positive information. In these areas, “negation as failure” assuming a closed world approach to reasoning is quite appropriate.

Most of the current trend concentrates on enhancing Semantic Web logics with reasoning, e.g. [12,4]. Efforts based on F-Logic [23] are used to provide a logic object-oriented basis of solutions to Semantic Web requirements. SWSL [13] is such an example for Semantic Web Services and is usable with XML-based standards, such as RuleML. Usually, most approaches combine rules with some form of description logic, e.g. [16,27]. A substantial effort is, currently, made to enhance Web with reasoning capabilities. The REWERSE Network of Excellence [31] on “Reasoning on the Web” (EU IST-FP6) was devoted to this purpose. As far as combining rules and ontologies, as needed in the Semantic Web, an overview of the encountered problems and how the proposed approaches address them can be found in [10].

During the last decades, it became obvious that a mechanism for “divide and conquer” is required in software construction. This was due to the fact that programs became larger. The proposed approaches fell under the general term “modularity”. The pioneer procedural language to make modules a first class concept was Modula-2 [41]. Modularity in logic programming has been thoroughly studied during late 80’s- early 90’s. A comprehensive review of the approaches can be found in [9]. Recently, as reasoning is on the focus on the Web and the Semantic Web communities, attempts to formalize reusable modules emerge. A generic approach for web rule languages is presented in [5]. This provides a basis for the module support of Xcerpt [42]. Xcerpt is a deductive rule-based query language able to operate both on traditional Web as well as Semantic Web environments.

In this work, we exploit methodologies based on modular logic programming that can also address the problems encountered in the Semantic Web. The proposal extends the work in [20]. In the next section, we present some background concepts on modularity by presenting two approaches that we are going to exploit. Next, we proceed with the addressed problems, starting from the information source level, information sharing, inheritance including multiple inheritance, reasoning and agents. The notion of subagents is, then, introduced and supported. Finally, we give our conclusions. Although we follow the Semantic Web architecture, we apply our methodologies to examples stemming from the traditional Web domain.

## 2 Modular Logic Programming

A programming language that provides modularity allows the programmer to develop each piece of code, namely “module”, separately, hiding (encapsulating) implementation details that are not to be used outside this module. The module

interchanges information with other modules via well defined interfaces. The module interfaces have to be fixed initially, as far as syntax and semantics are concerned. This has to be respected all the time on.

In logic programming, in general, there are two orthogonal directions:

1. the encapsulation affects only predicate symbols (*predicate based*) or all symbols (*atom based*), i.e. predicate symbols as well as function symbols and constants.
2. modularity is carried out by *program composition*, e.g. the approach in [6], or *linguistic extensions*, as in [26].

Concentrating, first, on predicate based modularity and, next, on program composition, in the following we present two composition operators and a flexible module system for logic programming. We claim that this background is suitable for modeling inheritance in taxonomies as well as agents for applications operating on both the traditional Web and the Semantic Web.

## 2.1 Two Composition Operators

*Union* is the simplest composition operator that can be applied on logic programs. It simply stands for the set-theoretic union of their clauses. For example, if  $P = \{q \leftarrow p, z\}$  and  $Q = \{q \leftarrow t, k\}$  then  $P \cup Q = \{q \leftarrow p, z, q \leftarrow t, k\}$ . In [7] the model-theoretic semantics of the union of two logic programs is given by the minimal model of the resulting program which is:

$$\begin{aligned} M_{P \cup Q} &= \cap \{M \mid M \models P \ \& \ M \models Q\} \\ &= \min \{M \mid M \models P \ \& \ M \models Q\} \end{aligned}$$

The union operator is monotonic.

On the other hand, given two logic programs  $P$  and  $Q$ , their *overriding union*,  $P \triangleleft Q$ , results to a program that contains the union of  $P$  and  $Q$  but if both  $P$  and  $Q$  contain a definition for the same predicate, then the definition of  $P$  overwrites the one found in  $Q$ . Considering the above programs  $P$  and  $Q$ , then  $P \triangleleft Q = \{q \leftarrow p, z\}$ . More formally, let  $\delta(P)$  be the set of predicate symbols defined by the program  $P$  and let  $Pred(A)$  denote the predicate symbol of any given atom  $A$ . Then  $P \triangleleft Q$  denotes the program obtained as follows:  $P \triangleleft Q = P \cup \{A \leftarrow G \in Q : Pred(A) \notin \delta(P)\}$ .

The semantics of  $\triangleleft$  is given in [8]. In this work, program *units* are introduced to denote individual pieces of code. The units of any program  $P$  are related through partial order  $\prec$ . We say that a unit  $u$  is a superunit of a unit  $u'$  if and only if  $u' \prec u$ . We call a unit  $u$  root unit if and only if there is no other unit  $u'$  such that  $u \prec u'$ . The inheritance relation between units is expressed in terms of the overriding union operator,  $\triangleleft$ , considered to be applied to units.

A SelfLog [8] program  $S = (U, \prec, R, \mu, \triangleleft)$  is built over a set of constants  $D$  and over a set  $\Pi$  of predicate symbols.  $U$  is a set of unit names,  $R$  is a finite set of extended rules,  $\mu$  is the mapping of rules to unit names,  $\triangleleft$  is the overriding algebraic operator that applies to units, and  $\prec$  is a partial order over  $U$  that defines a tree. The Herbrand base  $\mathcal{B}$  for each  $u \in U$  is built from  $D$  over  $\Pi$ . The semantics of SelfLog programs are given in [8].

## 2.2 A Module System for Logic Programming

In this paragraph, we present a flexible module system with import/export declarations that exchange predicate information at “extensional level”. That is to say, the encapsulation allows the logical consequences of a predicate to be visible outside the module rather than the predicate definition itself (the latter is referred to as the “intensional level”). The system and its formal semantics is presented in [21]. In the following, we give a brief outline of the framework.

A module encapsulates predicate names and definitions and is introduced by a module header, declaring its name. Then, the interface declarations follow that perform the predicate information exchange. More precisely, the interface declarations carry out the following. A module is allowed to declare all or some of its predicates *global* to the system, i.e. visible everywhere, “its predicates” referring to the ones defined in this module. In addition, it can declare them *exported* to specified modules or to all the modules of the system. Furthermore, a module may *import* predicates, at the extensional level, *from* specific modules. In this case, to achieve usability, the other module must export the predicates to the home module of the import declaration. Moreover, a module may *merge* procedure results *from* other modules, specific or not, with its own ones. Again an export declaration in the other module must export the procedure to the home module of the merge declaration.

The syntax of the module header and the interface declarations is:

**module header** *:-module*(*ModuleName*)  
**export declaration** *:-export\_to*(*ModName*,*Pred*)  
**import declaration** *:-import\_from*(*ModName*,*Pred*)  
**global declaration** *:-global*(*Pred*)  
**merge declaration** *:-merge\_from*(*ModNames*,*Pred*)

In order to determine which procedure definition is addressed when a predicate appears within a module, *visibility states* are introduced and define the effect of the interface declarations.

**Definition 1.** *A predicate can fall into one and only one of the following visibility states within a module:*

**merged** *iff a merge declaration exists in this module and refers to this predicate*  
**local** *iff the above does not hold and there is a definition for this predicate in this module*  
**imported** *iff none of the above holds and there is an import declaration in this module that refers to this predicate*  
**possibly\_global** *iff none of the above holds*

In the semantic framework, modules are considered as a kind of first order theories, namely *module First Order Theories (m-FOTs)*. Each module is a m-FOT. In [21] model theoretic, fixpoint and operational semantics were given as well as a transformation to Horn clause logic exhibiting that the framework is logical in the Horn clause sense.

### 3 The Addressed Problems

In the following, we consider the Semantic Web architecture, as proposed by Tim Berners-Lee [38]. Various issues arising in the Semantic Web layers, especially the lower ones, were examined taking a logic programming perspective in [20]. Briefly we can say here that, as far as the information source level is concerned, Web data can be considered as datalog knowledge bases [40]. This covers general n-ary relations without requiring to split information into binary relations (the RDF case).

#### 3.1 The Ontology Layer

More challenging issues reside in the ontology layer of the Semantic Web and up. It has to be noted, though, that in the traditional Web sources, in many cases, simple taxonomies rather than complex ontologies are required. This is the case when some entity organization and attribute assignment rather than general domain knowledge is required for the domain. Examples of such domains are, tourist [35], word news [19] or even news in general [11], computer science subject area [1], etc. A part of the taxonomy in [35] is shown in Figure 1.

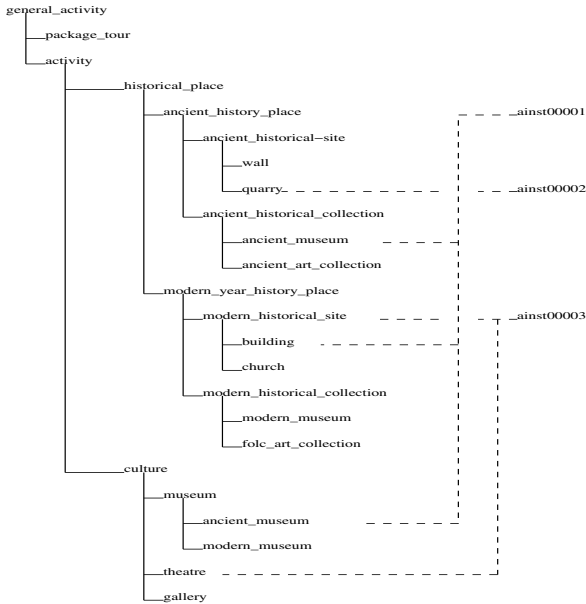


Fig. 1. A part of a touristical taxonomy and its instances

Taking into account such structures, logic programming representations can be employed in various ways. According to [20], one approach is to use Horn clauses to denote the concept subsumption and instantiation. In an alternative approach,



the concept relation is represented using facts. In case, though, attributes to concepts and instances are needed in the domain and if we need inheritance and default reasoning, results on inheritance in logic programming and datalog can provide the solution [8,2]. More precisely, concepts can be represented by classes or units in these frameworks. Subsumption can be represented by subclass relation and concept individuals can be represented by objects of classes. Both classes and instances are pieces of logic programs that are combined through inheritance. For the example of Figure 1, we can consider

```
class activity subclass general_activity
duration(60)
cost(0)
```

```
class historical_place subclass activity
```

```
class ancient_history_place subclass historical_place
time_period(01, 01, 31, 12)
closed_day(thursday)
```

```
class ancient_historical_collection subclass ancient_history_place
```

```
class ancient_museum subclass ancient_historical_collection
```

```
object ainst00001 of class ancient_museum
site(corfou)
denomination(ArcheologicalMuseumofCorfou)
closed_day(thursday)
closed_day(saturday)
descr(ArchaicsculpturesSilvercoins)
```

Monotone and non-monotone inheritance can be expressed in the logic program composition frameworks employing the *union* ( $\cup$ ) and *overriding union* ( $\triangleleft$ ) composition operators [9].

**Multiple Inheritance.** A problem that is often omitted when discussing about inheritance is multiple inheritance. In this case, a class is considered as a subclass of more than one classes and inherits information from all of them. Although in programming, multiple inheritance is a rare need and we have better avoid its usage, in knowledge representation it appears very often. Thus, multiple inheritance has to be discussed, introducing many problems, in case conflict exists between the information inherited from the superclasses.

In the approach that we follow, only positive information is considered. Then, the problem arises in case predicate name conflicts exist between different inheritance paths. What we propose is to declare a path preference order or even the specific path where a predicate is to be inherited from. In addition, we allow, if desired, to merge conflicting information from all paths rather than choosing a single path. Imposing preference on paths and choosing predicate from specific path is achieved by the non-monotone composition using overriding union.

The latter form, where all information is merged, is achieved by the monotone composition of union.

To clarify the above, consider a class  $C$  that is a direct subclass of classes  $C_1$ ,  $C_2$  and  $C_3$ . Consider that the preference is: first inherit from  $C_1$  and upward, then, if nothing found, from  $C_2$  and upward and, if nothing found, from  $C_3$  and upward. In our framework, we have to express this by declaring

$$\text{class } C \text{ subclass } C_1, C_2, C_3$$

To express the semantics of the above, we consider units  $U$ ,  $U_1$ ,  $U_2$  and  $U_3$  that correspond to the classes  $C$ ,  $C_1$ ,  $C_2$  and  $C_3$ , respectively. Then, the above subclass expression is modeled by

$$(U \triangleleft U_1) \triangleleft (U \triangleleft U_2) \triangleleft (U \triangleleft U_3)$$

In case, though, a predicate, e.g.  $p/n$ , is desired to be inherited by a specific path, e.g. the path starting from  $C_3$ , we need to declare this by

$$\text{class } C \text{ subclass } C_1, C_2, C_3 \text{ inherits } p/n \text{ from } C_3$$

We call the above *predicate inheritance specialization* and it is modeled as

$$(U \triangleleft [[U_3]]|p/n) \triangleleft (U \triangleleft U_1) \triangleleft (U \triangleleft U_2) \triangleleft (U \triangleleft U_3)$$

where  $[[U]]$  is the denotation of a unit  $U$ , as defined in [8], and  $[[U]]|p/n$  is its restriction on the predicate  $p/n$ .

Finally, if a predicate needs to be collected from more than one class, this can be declared by

$$\text{class } C \text{ subclass } C_1, C_2, C_3 \text{ inherits } p/n \text{ from } C_3 \text{ extends } q/m \text{ from } C_1, C_3$$

The above means

$$(U \cup [[U_1]]|q/m \cup [[U_3]]|q/m) \triangleleft (U \triangleleft [[U_3]]|p/n) \triangleleft (U \triangleleft U_1) \triangleleft (U \triangleleft U_2) \triangleleft (U \triangleleft U_3)$$

and is called *predicate inheritance extension*. Monotone inheritance is performed.

In the following, we give an outline of the procedure to build the unit expression:

1. In case of predicate extensions, combine their denotations by monotone union
2. Then, in case of predicate inheritance specialization, combine its denotation by overriding union
3. Then, add the inherited paths combined with overriding union
4. Finally, combine all the resulted unit expressions with overriding union

The above example has inheritance paths of length=1. If the length is longer, the procedure is recursive up to the base classes.

Considering the above tourist application, let's consider the following class (without its superclass)

```
class museum
closed_day(saturday)
cost(10)
```

Then, the *ancient\_museum* class could be defined as

```
ancient_museum subclass museum, ancient_historical_collection
extends closed_day/1 from museum, ancient_historical_collection
```

In this case, the *closed\_day/1* clauses need not appear in the *ainst00001* definition.

### 3.2 Reasoning and Agents

Having discussed data and metadata representation, we now continue by discussing reasoning and agents.

Distributed reasoning to cope with computational load has already been considered in other work, e.g. in [34]. Exploiting modularity and encapsulation, though, different sources may be accompanied with different inference rules, other than deduction, e.g. abduction [18] or induction [17]. We propose to consider the different information sources —accompanied with their inference rules— as modules. Then, the operational behaviour of the module system presented in [21] can be exploited.

In the following, we focus on agents and a specialization of them that we call “subagents”. Agents were part of Distributed Artificial Intelligence (DAI) that was revived in the context of the Web and Semantic Web. Agents are pieces of software that operate autonomously, often on behalf of some user or to provide information requested by other agents. Intelligence and mobility are often parts of their behavior.

Agents are considered to fall into different categories according to their characteristics and behavior [33]. Logic-based agents fit better to Semantic Web environment and the World Wide Web. The reason is that intelligence is of major importance both as far as behavior and knowledge management is concerned. For instance, in [39], abductive logic agents have been used to model information management in a World Wide Web environment. The dynamic nature of the agents' specification and reasoning environment is discussed in [3].

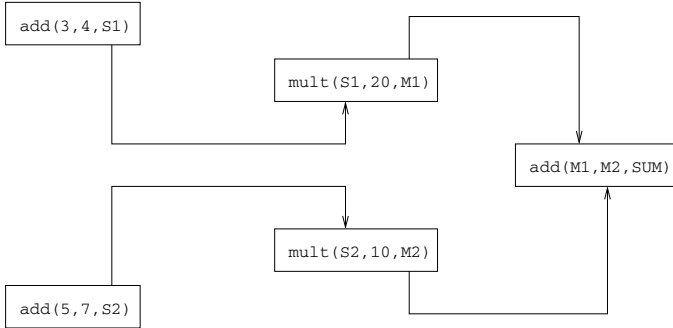
Our proposal was first introduced in [20]. More precisely, we considered agents to correspond to modules that exchange information using selective import-export declarations. Web site information sources can also be considered as modules where global declarations make all information available to the rest of the world.

**Subagents.** We can further refine the concept of agents into the concept of “subagents”. These are more lightweighted agents, private to an agent and can be used only by this agent. Subagents can be used to accomplish subtasks of the task that the agent has to carry out. By assigning subtasks of the original task to its subagents, the agents can complete their task more easily.

Let's consider a simple example. Let's have an agent that is able to compute mathematical formulae. This agent can work by invoking subagents, each for each mathematical operation, namely addition, subtraction, multiplication and division. Then, to perform the operation

$$(3 + 4) * 20 + (5 + 7) * 10$$

the agent activation presented in Figure 2 has to be carried out.



**Fig. 2.** Mathematic Operations Subagents

The general form of subagents is

```

:-subagent <subagent_name> / <arity> .
<subagent_code>
:-endsubagent <subagent_name> / <arity> .
  
```

providing just one predicate, appearing in its header.

Then, the code of subagents appearing in Figure 2, can be

```

:-subagent add/3.
add(X,Y,Z) :- all_integers([X,Y]), compute_result(X,Y,Z).
compute_result(X,Y,Z) :- Z is X + Y.
:-endsubagent add/3.

:-subagent mult/3.
mult(X,Y,Z) :- all_integers([X,Y]), compute_result(X,Y,Z).
compute_result(X,Y,Z) :- Z is X * Y.
:-endsubagent mult/3.
  
```

and the interface of the agent that performs the general computations can be

```

:-agent(computation_agent)
:-subagent(add/3).
:-subagent(mult/3).
  
```

The subagent concept can be modeled using the module system presented in Paragraph 2.2. More precisely, all agents and subagents are translated into individual

modules. Each subagent is translated into a module exporting its predicate only to its agent module. Each agent is also translated into a module that imports predicates from all its subagents' modules. To illustrate the mapping, the above agent interface as well as its subagents' are translated as follows:

```

: -module(computation_agent)
: -import_from(add3, add/3)
: -import_from(mult3, mult/3)
...
: -module(add3)
: -export_to(computation_agent, add/3)
...
: -module(mult3)
: -export_to(computation_agent, mult/3)
...

```

The subagent concept is important for a widely distributed environment such as the Web environment. The reason is that global information and functionality may conflict unpredictably because of the size of the domain space. So the more tools we have to keep information and/or functionality local, the more sound and predictable systems we construct.

## 4 Conclusions

To conclude, we discussed issues encountered in the World Wide Web and addressed in the Semantic Web effort and exhibited how they can be solved using Logic Programming. More precisely, we focused on Modular Logic Programming. Modularity can be exploited in the higher levels of the Semantic Web layers. Units and unit composition provides the means for subsumption representation of the ontology layer. The framework covers the case of multiple inheritance. Import/export modules can provide the means for distributed and even heterogeneous reasoning as well as agents. Heterogeneity is related with different inference rules that may accompany different information sources. Module organization covers the case of subagents, a concept introduced here to represent lightweight agents that perform auxiliary tasks that an agent needs.

Modularity has been also exploited in [22] to provide the logical basis for the interoperability between rules and ontologies. "Interoperability" is the alternative to "integration" as discussed, for instance, in [15]. In this paper, we claimed that there are cases, for instance many kinds of Web data applications, where Modular Logic Programming provides a uniform framework from the data layer to the higher layer of reasoning and agents.

In our future work, we intend to combine all the requirements into a single semantic logic programming framework based on logic programming composition and modularity. Within this framework, we intend to examine the possible forms of compositionality and integration that might be needed in the Web and the Semantic Web environment.

## References

1. The ACM computing classification system (1998), <http://www.acm.org/class/1998/>
2. Afrati, F., Karali, I., Mitakos, T.: On inheritance in object oriented datalog. In: Proc. of Int. Workshop on Issues and Applications of Database Technology, IADT 1998 (1998)
3. Alferes, J.J., Brogi, A., Leite, J.A., Pereira, L.M.: Logic programming for evolving agents. In: Klusch, M., Omicini, A., Ossowski, S., Laamanen, H. (eds.) CIA 2003. LNCS (LNAI), vol. 2782, pp. 281–297. Springer, Heidelberg (2003)
4. Antoniou, G., Bikakis, A.: A system for non-monotonic rules on the web. In: Antoniou, G., Boley, H. (eds.) RuleML 2004. LNCS, vol. 3323, pp. 23–36. Springer, Heidelberg (2004)
5. Assmann, U., Berger, S., Bry, F., Furche, T., Henriksson, J., Patranjan, P.L.: A generic module system for web rule languages: Divide and rule. In: Paschke, A., Biletskiy, Y. (eds.) RuleML 2007. LNCS, vol. 4824, pp. 63–77. Springer, Heidelberg (2007)
6. Brogi, A.: Program Composition in Computational Logic. Ph.D. thesis, Università di Pisa (1993)
7. Brogi, A., Mancarella, P., Pedreschi, D., Turini, F.: Composition operators for logic theories. In: Proc. of the Computational Logic Symposium (1990)
8. Bugliesi, M.: A declarative view of inheritance in logic programming. In: Joint International Conference and Symposium on Logic Programming. pp. 112–127 (1992)
9. Bugliesi, M., Lamma, E., Mello, P.: Modularity in logic programming. Journal of Logic Programming 19/20 (2004)
10. Eiter, T., Ianni, G., Krennwallner, T., Polleres, A.: Rules and ontologies for the semantic web. In: Baroglio, C., Bonatti, P.A., Małuszyński, J., Marchiori, M., Polleres, A., Schaffert, S. (eds.) Reasoning Web. LNCS, vol. 5224, pp. 1–53. Springer, Heidelberg (2008)
11. Fernandez-Garcia, N., Sanchez-Fernandez, L.: Building an ontology for NEWS applications. Poster Presentations at the International Semantic Web Conference ISWC-2004 (2004)
12. Grosz, B., Horrocks, I., Volz, R., Decker, S.: Description logic programs: Combining logic programs with description logic. In: Proc. of the 12th International World Wide Web Conference (WWW 2003). ACM, New York (2003)
13. Grosz, B., Kifer, M., Martin, D.: Rules in the semantic web services language (SWSL): An overview for standardization directions. Position paper for W3C Workshop on Rule Languages for Interoperability (2005)
14. Horrocks, I., Parsia, B., Patel-Schneider, P.F., Hendler, J.A.: Semantic web architecture: Stack or two towers? In: Fages, F., Soliman, S. (eds.) PPSWR 2005. LNCS, vol. 3703, pp. 37–41. Springer, Heidelberg (2005)
15. Horrocks, I., Parsia, B., Patel-Schneider, P., Hendler, J.: Semantic web architecture: Stack or two towers. In: Fages, F., Soliman, S. (eds.) PPSWR 2005. LNCS, vol. 3703, pp. 37–41. Springer, Heidelberg (2005)
16. Horrocks, I., Patel-Schneider, P.F.: A proposal for an OWL rules language. In: Proc. of the Thirteenth International World Wide Web Conference (WWW 2004). ACM, New York (2004)
17. Inductive logic programming, <http://www.doc.ic.ac.uk/~shm/ilp.html>
18. Kakas, A., Kowalski, R., Toni, F.: Abductive logic programming. Journal of Logic and Computation 2 (1993)

19. Kallipolitis, L., Karpis, V., Karali, I.: World news finder: How we cope without the semantic web. In: Proc. of Int. Conference on Artificial Intelligence and Applications (AIA 2007) (2007)
20. Karali, I.: Logic programming to address issues of the semantic web. In: 2007 IEEE / WIC / ACM International Conference on Web Intelligence. IEEE Computer Society, Los Alamitos (2007)
21. Karali, I., Halatsis, C.: A refinement of import/export declarations in modular logic programming and its semantics. In: Mosses, P.D., Schwartzbach, M.I., Nielsen, M. (eds.) CAAP 1995, FASE 1995, and TAPSOFT 1995. LNCS, vol. 915, Springer, Heidelberg (1995)
22. Kifer, M., de Bruijn, J., Boley, H., Fensel, D.: A realistic architecture for the semantic web. In: Adi, A., Stoutenburg, S., Tabet, S. (eds.) RuleML 2005. LNCS, vol. 3791, pp. 17–29. Springer, Heidelberg (2005)
23. Kifer, M., Lausen, G., Wu, J.: Logical foundations of object-oriented and frame-based languages. *J. ACM* 42(4) (1995)
24. Lloyd, J.W.: Foundations of Logic Programming. Springer, Heidelberg (1987)
25. Loke, S.W., Davison, A.: Logic programming with the world wide web. In: Proc. of the 7th ACM Conference on Hypertext. ACM, New York (1996)
26. Miller, D.: A theory of modules for logic programming. In: Proceedings of the 1986 Symposium on Logic Programming. pp. 106–114 (1986)
27. Motik, B., Horrocks, I., Rosati, R., Sattler, U.: Can OWL and logic programming live together happily ever after? In: Cruz, I., Decker, S., Allemang, D., Preist, C., Schwabe, D., Mika, P., Uschold, M., Aroyo, L.M. (eds.) ISWC 2006. LNCS, vol. 4273, pp. 501–514. Springer, Heidelberg (2006)
28. <http://www.w3.org/TR/owl-features/>
29. <http://www.w3.org/RDF/>
30. <http://www.w3.org/TR/rdf-schema/>
31. REVERSE: reasoning on the web with rules and semantics, <http://reverse.net/>
32. <http://www.ruleml.org/>
33. Sadri, F., Toni, F.: Computational logic and multi-agent systems: a roadmap. *ComputolNet Newsletter* (1999)
34. Schlicht, A., Stuckenschmidt, H.: Towards distributed ontology reasoning for the web. In: 2008 IEEE / WIC / ACM International Conference on Web Intelligence. IEEE Computer Society, Los Alamitos (2008)
35. Stamatopoulos, P., Karali, I., Halatsis, C.: A tour advisory system using a logic programming approach. *ACM Applied Computing Review* 1(1) (1993)
36. <http://www.w3.org/2001/sw/>
37. <http://www.w3.org/2003/Talks/0922-rsoc-tbl/>
38. <http://www.w3.org/DesignIssues/diagrams/sw-stack-2002.png>
39. Toni, F.: Automated information management via abductive logic agents. *Telematics and Informatics* (2001)
40. Ullman, J.: Principles of Database and Knowledge-Base Systems. Computer Science Press, Inc. (1988)
41. Wirth, N.: Programming in Modula-2. Springer, Heidelberg (1985)
42. Xcerpt-module system,  
<http://www.pms.ifi.lmu.de/reverse-wgi4/software/Xcerpt/modules>
43. <http://www.w3.org/XML/>

# Automatic Collecting Technique of Low Frequency Electromagnetic Signals and Its Application in Earthquake Study

Xuemin Zhang<sup>1,2</sup>, Roberto Battiston<sup>2</sup>, Xuhui Shen<sup>1</sup>, Zhima Zeren<sup>1</sup>, Xinyan Ouyang<sup>1</sup>, Jiadong Qian<sup>1</sup>, Jing Liu<sup>1</sup>, Jianping Huang<sup>1</sup>, and Yuanqing Miao<sup>1</sup>

<sup>1</sup> Institute of Earthquake Science, China Earthquake Administration, Beijing, 100036, China

zhangxm96@126.com

<sup>2</sup> Dipartimento di Fisica and °Sezione INFN, Via A. Pascoli, Perugia, Italy

**Abstract.** Two methods to automatically collect disturbed ultra low frequency electromagnetic signals before strong earthquakes were developed in this paper, in which one is only related to the spectrum intensity of electric field, and another is additionally combined with fractal dimension feature in electromagnetic emissions. Both techniques were applied for satellite data processing, case study and precursor distinction around Chile earthquakes, especially those occurring in 2010. Their advantages and disadvantages were compared and discussed based on actual application results. Further research on data mining needs to be carried out to improve the old and to develop the new ones based on the property of electromagnetic waves and multi parameters.

**Keywords:** Ultra Low Frequency electromagnetic emission; Fractal property; Chile earthquake; Multi parameters; DEMETER.

## 1 Introduction

Electromagnetic phenomena have been widely observed on surface and satellites, and they attract more and more attentions due to their short-term characteristics before earthquakes during recent 20 years [1,2]. The launch of DEMETER and other space projects reflects the global developing trend in seismo-electromagnetics [3].

Fraser-Smith et al.[4] first presented significant emissions of ULF magnetic field before Loma Prieta M7.1 earthquake in 1989, and many anomalous electromagnetic signals have been observed on the ground in Japan, Greek and other countries[5,6,7,8]. As for the satellite observation, a lot of statistical analysis has showed the correlation between electromagnetic field anomalies and strong earthquakes. An increase in the intensity of low-frequency (0.1-16kHz) radiowave emissions has been detected on Intercosmos-19 [9]. Malchanov et al.[10] summed up 28 earthquakes during Nov 16 1989 to Dec. 31, 1989 using Intercosmos-24 satellite data, and concluded that emissions with spectrum maxima occurred at two frequency bands, ULF-ELF ( $f$  less than 1000Hz) and VLF ( $f=10-15$ kHz) over the earthquake epicenter, and they were always detected 12-24 hours before the main shocks. Serebryakova



et al.[11] obtained similar EM radiation on COSMOS-1809 and AUREOL-3 with similar wave intensities and spectral distributions below 450Hz over the earthquake region in Armenia during Jan 20 to February 17, 1989.

All these illustrate the existence of electromagnetic signals before strong earthquakes and demonstrate a promising future of seismo-electromagnetics in earthquake study. However, there are few anomaly distinguishing methods in electromagnetic field at present. The main one is by looking through the spectral images, which is not scientific and relies on one's judgement sometimes; while another one is statistical method, which is not suitable for imminent anomaly extraction and distinction. Accompanied with the growing amount of observing data in recent years, it becomes a problem that how to select those disturbed electromagnetic signals automatically and reduce the man-made effects in data processing. In this paper, two methods were developed to pick up the EM emissions in satellite data, and the relationship between the anomalies and Chile earthquakes was discussed.

## 2 Data Processing Method

### 2.1 Method 1 - The Direct Selection of Electric Field Perturbations

Firstly, we give an example to show the shape of ultra low frequency electromagnetic disturbances observed on DEMETER satellite. On 14 Nov 2007, a M7.9 earthquake took place in Chile, and some significant anomalies were found prior to it [12]. As presented in Fig. 1, in the half orbit 17973 recorded 1 day before this Chile earthquake, ionospheric perturbations were observed at  $\pm 20^\circ$  of latitude in most parameters. From the top panel to the last, Fig.1 shows the DC-1000Hz electric field spectrogram (E), 19.5-1000Hz magnetic field spectrogram (B), electron density (Ne), electron temperature (Te), total ion density (Ni), ion density (Ni, H+, He+, O+), ion temperature (Ti) and earthquakes distribution in  $\pm 30$  days around this orbit in a distance of 2000km. It can be seen clearly that Ti varied significantly, rising more than 1000K. Simultaneously, other parameters were disturbed to some certain extent. Here we focus on the selection of the disturbances lower than 250Hz in very low frequency (VLF) electric field spectrograms. After repetitive learning and imitating process on many similar signals, it is found that the EM perturbations are frequently with spectrum values larger than  $10^0 \mu V^2 / m^2 / Hz$  (sometimes much higher than that) at the first frequency points, reduce subsequently, and then degrade to the same level with the background after 250Hz that means the perturbations end at that frequency. So the first method is defined directly based on this characteristic in ULF electromagnetic perturbations. To allow a convenient comparison and visualized figures, the points where electric field intensities are larger than  $10^0 \mu V^2 \cdot m^{-2} \cdot Hz^{-1}$  at first and damp gradually to the normal from 19.5 to 250Hz are chosen and assigned values of 1, and those observing points with spectrums lower than  $10^0 \mu V^2 \cdot m^{-2} \cdot Hz^{-1}$  at first

frequency band were assigned as values of 0. In addition, those points with high intensity but no decayed feature were also made as 0. Therefore, the perturbations along the orbits at ULF band could be picked up by this technique only if its ULF spectrums accord with the condition referred above.

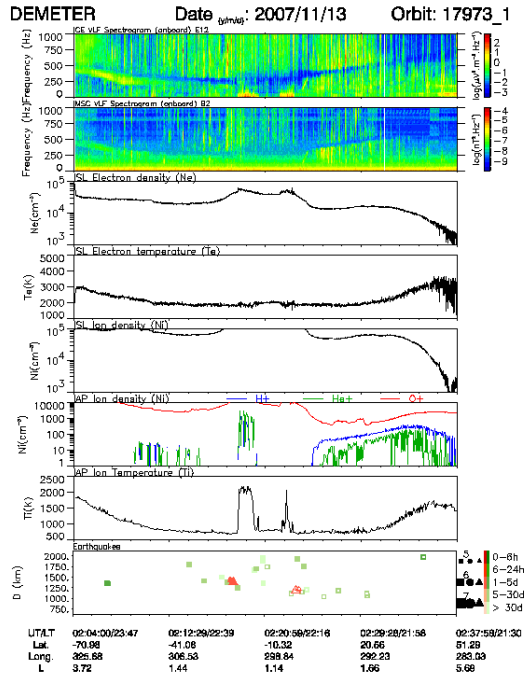


Fig. 1. The level-2 picture recorded at DEMETER satellite on 13 Nov. 2007

### 2.2 Method 2- The C Function Method

In method 1, only the external features of electromagnetic signals were considered, and no property of electromagnetic waves was taken into account. So maybe many other undesired signals will be mixed into the anomaly selection at the same time, which will place more weight on anomaly distinguishing in earthquake study. During DC-250Hz band, only the VLF spectrum data was available on DEMETER, because ELF electromagnetic waveforms lower than 125Hz were only recorded at BURST mode in some pre-designed seismic regions, not on whole orbit and the ULF electric field waveforms along the whole orbit is under the maximum frequency of 20Hz. In previous studies, the electromagnetic signals have shown the fractal feature in some ground-based ULF magnetic field data processing [13, 14]. According to their description, we can get the equation (1) to define the exponential relation between the spectrum data ( $S_E$ ) and frequency ( $f$ ).

$$S_E = a \cdot f^{-b} \tag{1}$$

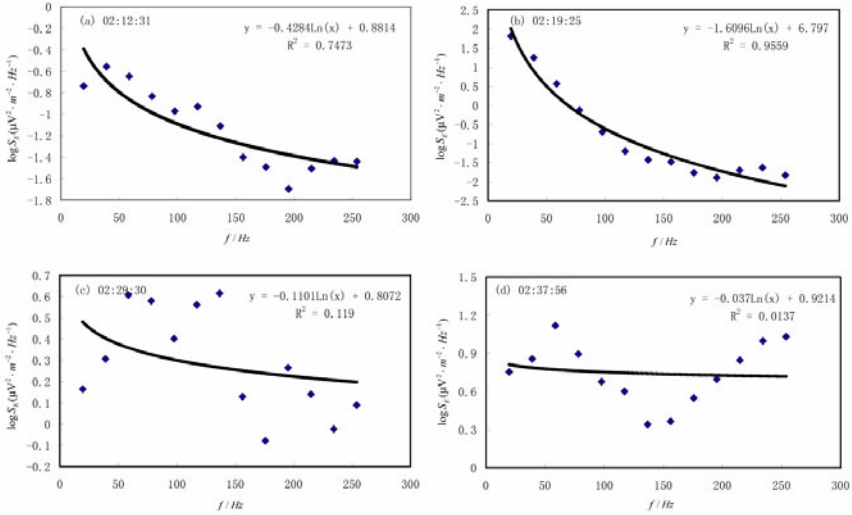


Fig. 2. Fractal parameter fitting at different observing points

Here 17973-1 orbit shown in Fig. 1 was taken as an example again. The spectral data at 19.5-250Hz in electric field were chosen at 4 different observing points (the time was shown after the (a)-(d) at top left corner in each panel of Fig. 2). Compared with the recording time in Fig. 1, Fig 2a shows a point located at southern hemisphere, whose electric spectrums (the blue diamonds in Fig.2) are all lower than  $10^0 \mu V^2 \cdot m^{-2} \cdot Hz^{-1}$ . But the exponential fitting (the black lines in Fig. 2) is quiet good with the correlation coefficients about 0.86. Fig 2b shows a point just located in the electric field perturbations near the equator. It shows well fractal feature, the fitting correlation being nearly 0.98. Fig 2c is located at mid latitude of north hemisphere. Although most of the spectral values are bigger than  $10^0 \mu V^2 \cdot m^{-2} \cdot Hz^{-1}$ , the signals does not show any scaling characteristic, just a noisy type. Fig 2d is the last point of 17973-1 orbit at high latitude above  $50^\circ N$ . The spectral values are relatively higher, but the fitting correlation coefficient is lower than 0.12. Among all the 4 points, only the point within electromagnetic disturbances exhibits the highest values in fractal dimension feature of parameter-b, and correlation coefficient-R.

On the basis of fitting results in Fig 2, it can be concluded that if one wants to distinguish the electromagnetic emissions, 3 factors need to be considered generally, the exponential parameter-b, the correlation-R and the initial intensity related to fitting parameter  $a$  in equation (1). Equation (2) is constructed by taking the logarithm and the absolute value to the right side of equation (1), in which  $a_0 = \lg a$ . We define it as Function  $C$ .

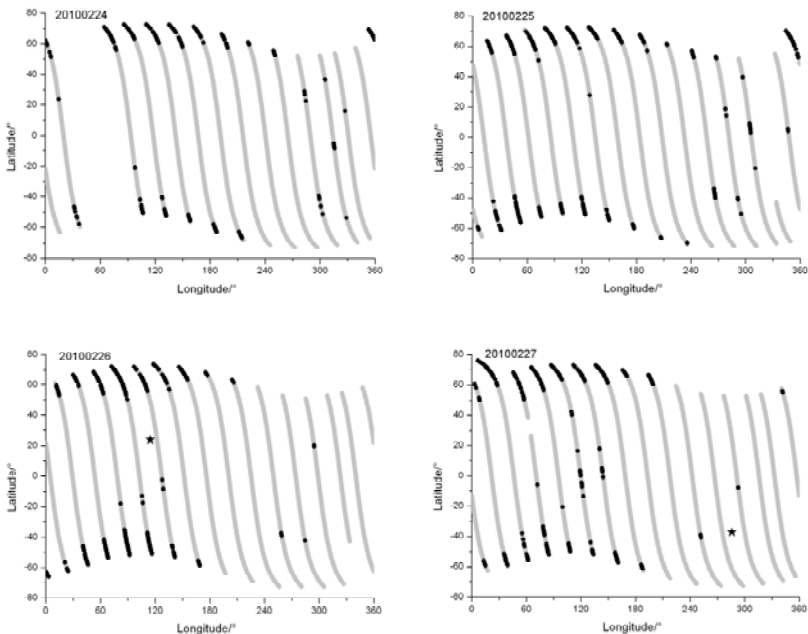
$$C = a_0 + b \times |R| \quad (2)$$

### 3 Application in Chile Earthquake Study

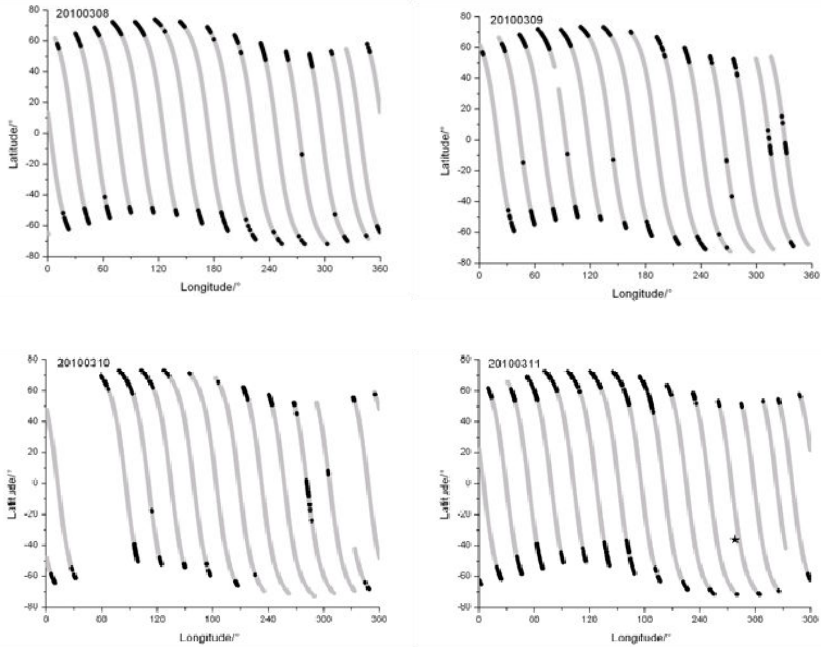
On 27 March 2010, a M8.8 earthquake took place in Chile, with the location of 35.93°S, 72.78°W (<http://neic.usgs.gov>). This event is the second strongest earthquake in this century in the world after the Sumatra M9.0 earthquake on 26 Dec. 2004 in Indonesia. Furthermore, on 11 March 2010, two strong aftershocks with  $M \approx 6.5$  occurred at same area. Following we deal with the DEMETER satellite data in electric field spectrums around this Chile earthquake sequence.

#### 3.1 The Signal Collection by Method 1

Employed Method 1, the disturbances in local nighttime in the VLF electric field were chosen and plotted in Fig. 3 during February 24-27 around the Chile M8.8 earthquake. The black dots in Fig 3 represent the value 1, and gray dots represent value 0, which means that every black dot is an intensive electromagnetic emission. It shows that electric perturbations exist widely at higher latitudes around  $\pm 60^\circ$ , especially in eastern hemisphere. Besides that, there are some signals located at mid and low latitudes, which occurred occasionally and may be associated with earthquakes. During February 24-27, these selected anomalies at low latitude in each panel in Fig. 3 are mainly distributed in two regions, 300-360°E, and 100-150°E. On 24, 25 and 27, these signals located at 300-360°E should be related to Chile 8.8 earthquake (black star on the image of the date-20100227) near this region. On 26 and 27, the signals at 100-150°E may be associated with another M7.0 earthquake (black star on the image



**Fig. 3.** The electric perturbations selected by method 1 during Feb. 24 to 27 2010



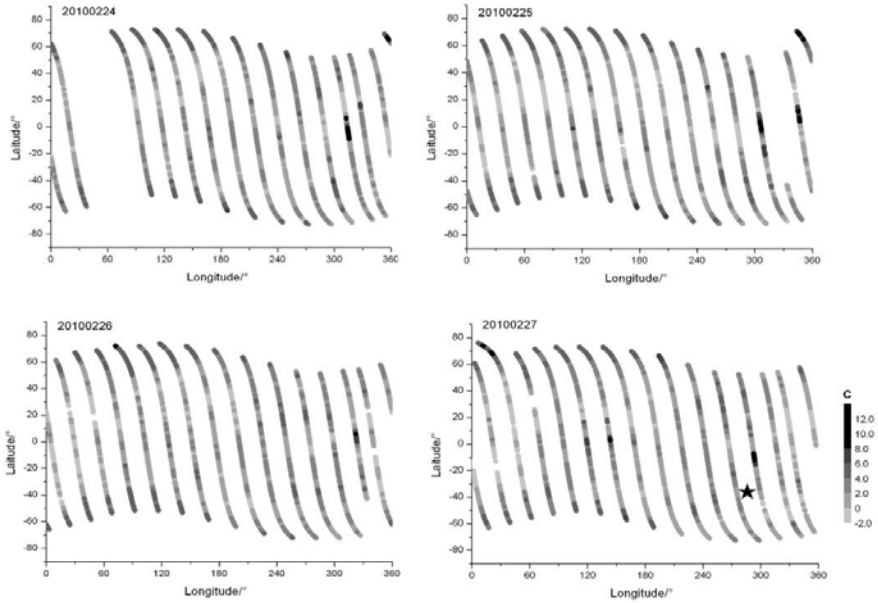
**Fig. 4.** The electric perturbations selected by method 1 during Mar. 8 to 11 2010

of 20100226) in the Ryukyu islands on 26 February 2010, 25.9° N, 128.42°E, where many anomalies occurred at its conjugate point and the after-seismic ionospheric effects were apparent over this seismic region and the area to the south.

Fig. 4 shows the automatic signal selecting results during 8 to 11 March 2010 around the strong Chile aftershocks. It reflects that on March 9 and 10, 1 and 2 days just prior to the strong aftershocks, the electric field perturbations occurred over the region of 270-360°E. And on March 10, these anomalous signals were only recorded on the nearest orbit over the area to the north of the aftershocks (see black star on the last panel), which illustrates close spatial relationship between these emissions and earthquakes in Chile. On 11, there was no obvious signal found over the epicentral area when the aftershocks were taking place.

### 3.2 The Signal Collection by Method 2

Using Function C, the global up-orbit VLF electric field data (19.5-250Hz) in local nighttime was processed during February 24-27 and March 8-11 in 2010 (Fig. 5 and 6). Different with Fig 3, anomalies with  $C \geq 8.0$  in Fig. 5 only occurred at 280-360°E region, and were shown every day during 4 days before Chile M8.8 earthquake. The signals at higher latitudes almost disappear after the calculation of C Function, which reflects that the EM emissions at high latitudes have different electromagnetic properties with those at low latitudes possibly associated with earthquakes. During 24<sup>th</sup> to

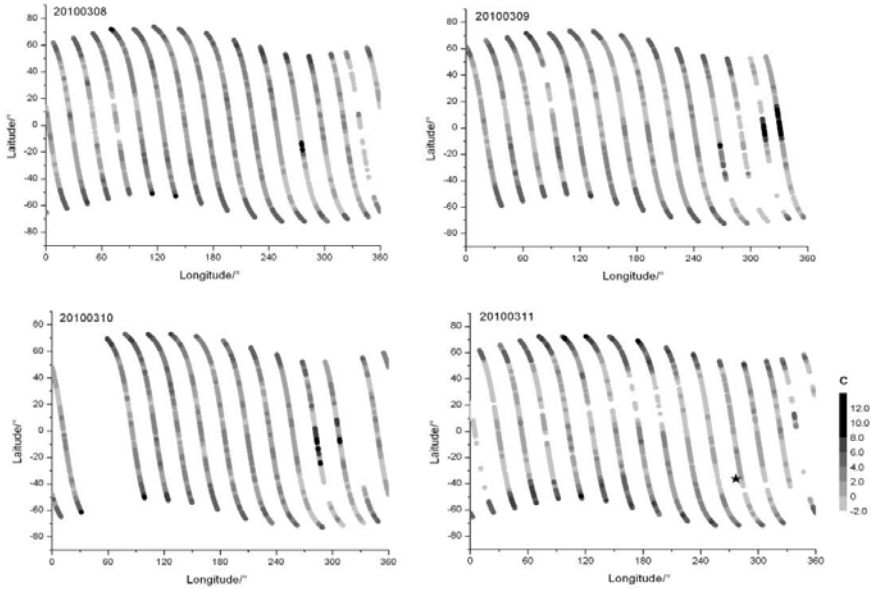


**Fig. 5.** The electric perturbations selected by method 2 during Feb. 24 to 27 2010

26<sup>th</sup>, the emissions mainly distributed to the northeast of the epicenter (the black star on the last panel). However, on 27<sup>th</sup>, the day when Chile M8.8 earthquake took place, the electromagnetic perturbations were recorded on the closest orbit from the epicenter, with the maximum C values in these 4 days.

The images of C Function during March 8 to 11 were exhibited in Fig. 6. The anomalous signals in electric field were mainly located to the far east of the epicenters on March 9, while on 8 and 10 they distributed over the region from 0° to 20°S at nearer orbits, which are similar with those in Fig. 4. On 11<sup>th</sup>, no obvious electric field emissions were obtained.

Compared with Method 1, the anomalous signals in electric field collected by Method 2 are much more reliable, because many emissions in Fig 3 and 4 at high latitudes disappear in Fig 5 and 6 but those at mid and low latitudes are kept as the same, which will largely reduce the possibility of misjudgment in precursor distinguishing. Moreover, Fig 5 shows much more emissions with high C values near the Chile earthquake epicenter on 27 February, which will play important role in predicting the location of future earthquakes. Undeniably, the Method 1 has its own advantages, because it trustily reflects the enhanced electromagnetic waves in low-frequency band, not concerning what they are resulted from, so it will not lose any similar signals. We may need to do further study on this method, such as dividing the anomaly type and doing cluster analysis and so on to improve its efficiency.



**Fig. 6.** The electric perturbations selected by method 2 during Mar. 8 to 11 2010

### 3.3 The Features of Ionospheric Anomalies in Chile Area

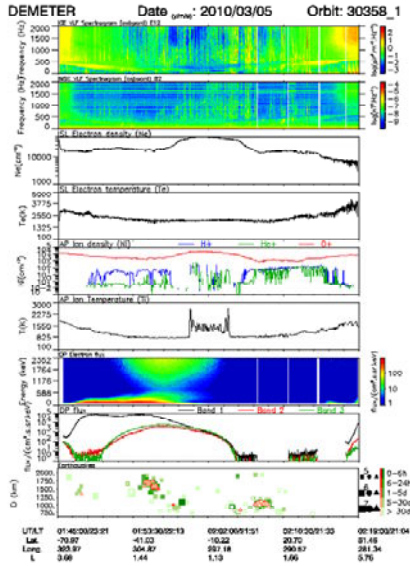
Chile is located in the Circum-Pacific seismic belt, a high frequent earthquake area due to the subduction of plates. Do all the Chile earthquakes have similar anomalies as those above? Following we show the statistical results about it.

Since June 2004, there took place 14 earthquakes with  $M \geq 6.5$ , which were divided into 10 groups according to their occurrence time (listed in Table 1). We selected the local nighttime data recorded on DEMETER satellite in 5 days around these earthquakes with 3 days before, the earthquake day and 1 day after. By using the method 1 and 2, it is found that there are 8 groups of earthquakes in 10 which showed ULF electric field anomalies, and simultaneous multi-parameter perturbations prior to them. The ionospheric EM disturbances always occurred over the epicentral area at first, and then moved to the equatorial area 1 day or a few hours before the main shocks while accompanying with more intensive amplitude, just as that presented before Chile M7.9 earthquake [12] and cases above in 2010. Only around two earthquakes on 25 Aug. 2006 and 5 March 2010, no apparent disturbances were obtained in electric field. But it is interesting that ion temperature ( $T_i$ ) displayed step variations before or after both of them. Fig. 7 gives the picture recorded on 5 March 2010, 9 hours prior to a M6.6 aftershock. The panels from the top represent respectively: DC-2000Hz electric field spectrogram (E), 19.5-2000Hz magnetic field spectrogram (B), electron density ( $N_e$ ), electron temperature ( $T_e$ ), ion density ( $N_i$ ,  $H^+$ ,  $He^+$ ,  $O^+$ ), ion temperature ( $T_i$ ), 70keV-2.3MeV electron flux, counting rate at 3 energy bands and earthquakes distribution in  $\pm 30$  days around this orbit in a distance of 2000km. It can

be seen that only Ti was disturbed a lot in the equatorial area, while other parameters maintained smooth curves and normal images over the studied area. It raises another question, how to comprehensively consider the multi-parameter anomaly and those only occurring in one parameter.

**Table 1.** The catalogue of Chile earthquakes with  $M \geq 6.5$  since 2005 (<http://neic.usgs.gov>)

CAT	Year	Month	Day	Time	Latitude	Longitude	Depth	Magnitude	Perturbed parameters
PDE	2004	8	28	134125.60	-35.17	-70.53	5	6.5	E,Ne,Ni,Ti
PDE	2005	11	17	192654.49	-22.36	-67.89	147	6.8	E,Ne,Ni,Ti
PDE	2006	4	30	191714.98	-27.02	-71.02	12	6.7	E,Ne,Ni,Ti
PDE	2006	4	30	214058.44	-27.21	-71.06	12	6.5	E,Ne,Ni,Ti
PDE	2006	8	25	4446.16	-24.4	-67.03	184	6.6	Ti
PDE	2007	11	14	154050.53	-22.25	-69.89	40	7.7	E,Ne,Ni,Ti
PDE	2007	11	15	150558.35	-22.92	-70.24	26	6.8	E,Ne,Ni,Ti
PDE	2007	12	16	80917.93	-22.95	-70.18	45	6.7	E,Ne,Ni,Ti
PDE-Q	2010	2	27	63414.23	-35.93	-72.78	35	8.8	E,Ne,Ni,Ti
PDE-Q	2010	2	27	80123.55	-37.75	-75.1	37	6.9	E,Ne,Ni,Ti
PDE-Q	2010	3	5	114707.14	-36.6	-73.23	18	6.6	Ti
PDE-Q	2010	3	11	143944.15	-34.26	-71.93	11	6.9	E,Ne,Ni,Ti
PDE-Q	2010	3	11	145527.97	-34.28	-71.84	18	6.7	E,Ne,Ni,Ti
PDE-Q	2010	3	16	22157.93	-36.22	-73.25	18	6.7	E,Ne,Ni,Ti



**Fig. 7.** The level-2 picture recorded on DEMETER on 5 March 2010



## 4 Results and Discussion

As presented in Fig. 3 and Fig. 4, the electric field perturbations selected by Method 1 are easy to be figured out, but some signals with weak spectral values may be lost such as those on 26 and 27 February 2010 over Chile seismic region. Moreover, many emissions are included at high latitudes possibly related to polar current activities. While the disturbed signals calculated by Function C can well reflect the EM emissions with certain intensities and fractal dimension feature. It shows special advantage in precursor selection of earthquake study. It is thought that more physical properties of electromagnetic waves such as the propagation and polarization features [15] may help to collect more reliable signals associated with earthquakes in future.

In this paper, the Chile earthquake sequences were analyzed by both methods. The results have illustrated the validity of two automatic techniques in VLF electric field spectrum data and high efficiency in spatial comparison of global orbits. Summing up all the  $M \geq 6.5$  earthquakes in Chile since 2004, the ionospheric perturbations before them can be divided into two kinds: one is with disturbed multi parameters consisting of electric field; and one is only in ion temperature. Based on this feature, some new ideas may be considered in next work. In equation (2), only electric field was included although the fractal feature has been taken into account. Maybe more items should be appended in future, such as Ne, Ti, and the correlation between different parameters, and so on. As for the sole Ti variation, the Method 1 can be replaced by the relative varying amplitude of Ti. It is well-known that, the ionospheric precursors related to earthquakes are in complex and changeable forms, sometimes to increase, sometimes to decrease, and different parameters exhibit their individual shapes, sometimes to be coincided with each other, sometimes to be inverse, which resulted in the difficulty of precursor distinction. Actually, the selection of anomalies in this paper is just the first step in earthquake study, and more data mining methods should be introduced into further research.

On the basis of case studies, we developed two methods to automatically select ULF electric field perturbations in satellite data processing, and they actually help to strengthen the scientificity and improve the work efficiency in actual application. The main conclusions can be drawn as follows about the methods and Chile earthquakes.

(1) Two methods have their advantages and disadvantages respectively, and method 2 gets more signals related to earthquakes. More physical property in electromagnetic emissions should be considered in future to enhance the ability of precursor distinction.

(2) There are 8 in 10 Chile earthquakes showing electric field anomalies accompanying with other parameters. It demonstrates the correlation between ionospheric perturbations and Chile earthquake.

(3) The electric field precursors firstly distributed at the circumjacent area over Chile seismic region, and then occurred near the epicenters with the maximum function C 1 day or a few hours before Chile earthquakes. It reflects the regional property of electromagnetic emissions associated with earthquakes due to their close spatial distribution.

(4) Some precursors of Chile earthquakes occurred in multi parameters, some only in one. More data mining methods should be introduced to discriminate the complex forms of ionospheric anomalies in satellite data.

## Acknowledgements

This paper is funded by the Basic Research Fund of Institute of Earthquake Science, CEA (02092408), International Cooperation Project (2009DFA21480) and the National Science and Technology Support Project (2008BAC35B01, 2008BAC35B05). We are grateful to the DEMETER Data Centre for provision of the satellite data. The Author Xuemin Zhang thanks Chinese Scholarship Council and CEA for supporting her study in INFN, Italy.

## References

1. Pulnits, S.A., Boyarchuk, K.A.: *Ionospheric Precursors of Earthquakes*, pp. 1–287. Springer, Berlin (2004)
2. Zhang, X., Zhao, G.Z., Chen, X.B., Ma, W.: Seismo-electromagnetic observation abroad. *Progress in Geophysics* 22(3), 687–694 (2007) (in Chinese with English Abstract)
3. Lagoutte, D., Brochot, J.Y., de Carvalho, D., et al.: The DEMETER science mission centre. *Planetary and Space Science* 54(5), 428–440 (2006)
4. Fraser-Smith, A.C., Bernardi, A., McGill, P.R., Ladd, M.E., Helliwell, R.A., Villard, O.G.: Low-frequency magnetic field measurements near the epicenter of the Ms 7.1 Loma Prieta Earthquake. *Geophysical research Letters* 17(9), 1465–1468 (1990)
5. Hayakawa, M., Ito, T., Hattori, K., Yumoto, K.: ULF electromagnetic precursors for an earthquake at Biak, Indonesia on 17 February 1996. *Geophys. Res. Lett.* 27, 1531–1534 (2000)
6. Hattori, K., Akinaga, Y., Hayakawa, M., Yumoto, K., Nagao, T., Uyeda, S.: ULF magnetic anomaly preceding the 1997 Kagoshima Earthquakes. In: Hayakawa, M., Molchanov, O.A. (eds.) *Seismo Electromagnetics Lithosphere–Atmosphere–Ionosphere Coupling*, pp. 19–28 (2002)
7. Varotsos, P., Alexopoulos, K.: Physical properties of the variations of the electric field of the earth preceding earthquakes (I). *Tectonophysics* 110, 73–98 (1984a)
8. Varotsos, P., Alexopoulos, K.: Physical properties of the electric field of the earth preceding earthquakes (II): Determination of epicenter and magnitude. *Tectonophysics* 110, 99–125 (1984b)
9. Larkina, V.I., Migulin, V.V., Molchanov, O.A., Kharkov, I.P., Inchin, A.S., Schvetcova, V.B.: Some statistical results on very low frequency radiowave emissions in the upper ionosphere over earthquake zones. *Physics of the Earth and Planetary Interiors* 57(1–2), 100–109 (1989)
10. Molchanov, O.A., Mazhaeva, O.A., Goliavin, A.N., et al.: Observation by the Intercosmos-24 satellite of ELF-VLF electromagnetic emissions associated with earthquakes. *Ann. Geophysicae* 11, 431–440 (1993)
11. Serebryakova, O.N., Bilichenko, S.V., Chmyrev, V.M., Parrot, M., Rauch, J.L., Lefeuvre, F., Pokhotelov, O.A.: Electromagnetic ELF radiation from earthquake regions as observed by low-altitude satellites. *Geophys. Res. Lett.* 19(2), 91–94 (1992)
12. Zhang, X., Qian, J., Ouyang, X., et al.: Ionospheric electromagnetic perturbations observed on DEMETER satellite before Chile M7.9 earthquake. *Earthq. Sci.* 22, 251–255 (2009)
13. Hayakawa, M., Molchanov, O.A., Biagi, P., Vallianatos, F. (eds.): *Seismo Electromagnetics and Related Phenomena*. Special issue of *Physics and Chemistry of the Earth* 29, 4–9 (2004)

14. Simirnova, N., Hayakawa, M., Gotoh, K., Volobuev, D.: Scaling characteristics of ULF geomagnetic fields at the Guam seismoactive area and their dynamics in relation to the earthquake. *Natural Hazards and Earth System Sciences* 1, 119–126 (2001)
15. Santolík, O., Nemeč, F., Parrot, M., Lagoutte, D., Madrias, L.: Analysis methods for multi-component wave measurements on board the DEMETER spacecraft. *Planetary and Space Science* 54(5), 512–527 (2006)

# Improving Search in Tag-Based Systems with Automatically Extracted Keywords

Ruba Awawdeh and Terry Anderson

School of Computing and Mathematics, University of Ulster, Newtownabbey  
BT37 0QB, Northern Ireland  
awawdeh-r@email.ulster.ac.uk, tj.anderson@ulster.ac.uk

**Abstract.** Tag-based systems are used by millions of web users to tag, save and share items. User-defined tags, however, are so variable in quality that searching on these tags alone is unsatisfactory. One way to improve search in book-marking systems is by adding more metadata to the user-created tags to enhance tag quality. The additional metadata we have used is based on document content and largely avoids the idiosyncratic and ambiguous terms too often evident in user-created tags. Such an approach adds value by incorporating information about the content of the resource while retaining the original user-created tags.

This paper describes how users' tags can be enhanced with metadata automatically extracted from the original document. An experiment comparing search based only on user-created tags with search using an automatically enhanced tag set, demonstrates how incorporating the extra tags can offer significant benefits.

**Keywords:** Tagging, Folksonomy, Searching.

## 1 Introduction

Tag-based systems are a major part of the interactive, collaborative trend in web software known as Web 2.0, and their popularity reflects their usability among web users. Tag-based systems, or 'folksonomies', add another dimension to web browsing and searching by drawing on social participation. These systems are distributed informal classification systems created by users who insert selected web resources into an online database and freely add unconstrained keywords to describe them. These resources can then be shared among other users or kept private. Sharing resources allows users with similar interests to discover links, documents and other kinds of online items [9], which can lead to a collection of more or less closely related items that may not be reported by conventional search engines.

Despite the huge success of folksonomies, these systems still face some problems which are limiting their effectiveness. One of the main problems is tag ambiguity. For example, a word may have multiple meanings ('polysemy'), so 'table' may refer to a piece of furniture or a grid of information. The converse ('synonymy') may be illustrated by the words 'code' and 'program', where several words refer to the same concept. Also, because users are free to add any tags they wish, some tags are arguably not even true metadata, as would be the case where they reflect personal

views or planned uses e.g. ('interesting', 'to-read') [1,6]. Without a controlled vocabulary or even a widely accepted set of guidelines, folksonomies exhibit limitations when searching and re-finding items [15].

Most of the websites that employ folksonomies are not designed to act as search engines and information retrieval is not their main aim. However information retrieval is a very important aspect in any website containing a large collection of resources, such as bookmarks or images. Therefore additional methods to improve searching and to enhance the user-created tags are needed to improve their value in searching tag-based databases.

This paper is structured as follows. In the next section the research motivation is given. Section 3 provides a brief literature review related to research in the area of search in tag-based systems, and Section 4 describes the prototype and methodology used in this research to enhance user tags with terms automatically extracted from the original documents using the Yahoo term extraction service. A brief description of Enhanced Tag Search engine (ETS) is also given. Section 5 reports on the experiment used to evaluate the ETS using different sets of tags and the results are discussed in Section 6. Finally, the conclusions are presented.

## 2 Motivation

Since their origins around 2002/2003, tag-based systems have multiplied, including sites such as Digg, Del.icio.us, Flickr and Connotea. While the freedom users have in creating their own tags is a major reason behind their success, the ambiguity in tags, including non-standard abbreviations, misspellings, polysemy and synonymy reduces the value of tags in searching a folksonomy [9]. Also, the number of tags a user attaches to an item is highly variable. In an analysis of 60,000 tagged items on delicious we found that the number of tags per item ranged from 0 to 19, though the modal number was just two. It is therefore unsurprising that searching based on user tags is often disappointing. From this, it is logical to argue that user-created tags need to be enhanced in some way to improve their value in searching a folksonomy-based site [3]. Our aim in this study is to improve search in folksonomies. This will involve adding context related metadata extracted from the original documents.

## 3 Related Studies

A number of studies are presented in this section in order to give a better understanding of the deficiencies of user-created tags and how these may be addressed. Starting with studies that have attempted to assess the value of user tags in search, this section will also include work on the use of controlled vocabularies and clustering techniques. Then research on the automatic extraction of tags is presented, and lastly the problem of ranking query results generated by tag-based search systems is described.

Golder and Huberman [8] analysed tag-based systems to gain a better understanding of the nature of user-created tags. They studied a sample of tags extracted from

del.icio.us, and found that some tags provide meta keywords (e.g. 'interesting', 'to-read') which indicate a personal opinion or intention without reflecting the content of the document. Nevertheless, they considered that most of the personal tags were useful to other users, broadly corresponding with research carried out by Michlmayr [14] whose study on tag properties concluded that whilst search based purely on user-created tags is of limited success it could be of value in filtering results.

The limitations of tag-based systems have also been documented by Krause et al. [11] who compared searching in these systems with traditional web search methods by studying the end user behaviour and ranking methods used in both systems. They used two datasets in their comparison. The first dataset was extracted from del.icio.us and the other was extracted using traditional search engines (MSN, Google and AOL). User behaviour was captured in both systems and they found that most users of tag-based systems focused on just a few topics, mainly IT, unlike users of major web search engines.

Kipp [12] studied tag-based systems from a different angle, where he compared the search process between a tag-based system (CiteULike) and information retrieval using a controlled vocabulary provided by Pubmed, an online medical journal database. The comparison was carried out to determine whether participants found tags useful in information retrieval or not. In the experiment searchers were asked to search for articles on a specific topic on both sites. While participants were able to use the controlled vocabulary in Pubmed, they preferred the tag-based system. Nevertheless Kipp states that a further study is required to confirm this finding.

In an effort to address some limitations of tag-based systems, Guy and Tonkin [9] suggested an approach to improve the quality of user-created tags by "educating users to add 'better' tags and improving the systems to allow 'better' tags to be added". Similarly Yang [16] suggested the use of a controlled vocabulary as a solution to reduce tag ambiguity by providing a list of similar words for each tag. Macgregor and McCulloch [13] also recommend training users and encouraging them to follow definite tagging-based patterns to improve the quality of tags as well. However, as yet there is no user-acceptable way to enforce this and it could be argued that the prospect of any such impositions runs counter to the spirit of the internet.

An alternative approach to improving tag-based systems is to augment user-created tags through the use of clustering algorithms in order to improve the search capability of such systems. Begelman et al. [5] demonstrate a clustering algorithm to be used in tag-based systems. Their algorithm is based on counting the co-occurrences between tags (tags that are used to describe the same link) and determining a cutoff point to decide when the degree of co-occurrence should be considered significant. Strongly related tags are identified once an appropriate cutoff point is established. An undirected graph is then constructed consisting of nodes, representing the tags, which are connected by weighted edges computed on the basis of tag co-occurrence. An edge will exist between 2 nodes only if they are strongly related. Each set of connected tags forms a cluster. The experimental results showed that related tags can be used to suggest tags for users while searching, exploring or even tagging.

By contrast, Brooks and Montanez [7] adopted a different approach comparing the effectiveness of two methods for clustering and naming blogs. Direct use of user-created tags was the basis of the first approach, while the second involved automatic extraction of words from the blog articles where the three words with the top TF/IDF (term frequency – inverse document frequency) were extracted. Their results illustrate that user-created tags did help in clustering the blogs, although they were poor for describing the overall subject matter of an article, a task much more successfully achieved using automatically extracted words.

Al-Khalifa [2] has performed a comparison between user-created tags and automatically extracted words from the Yahoo API term extraction tool. While her experiment demonstrates the value of these tags when used as annotations, she does not demonstrate how they can be used as searchable keywords [2]. Her results were particularly influenced by the scientific nature of the information in her dataset. The users were professionals who strongly tended to use well-recognized scientific terms as tags. The terms extracted by the Yahoo API service were less satisfactory for this dataset in that it did not reliably return the scientific terms on which it was so heavily reliant. The Yahoo API service has been widely used and proved beneficial in extracting meaningful terms, as has been documented on the Yahoo Developer Network. Evidence of its value comes from its widespread use and forceful reaction from users against plans to discontinue the service in addition to subsequent attempts to emulate the service e.g. Zemanta API. [19]

Bao et al.[4] proposed two novel algorithms, SocialSimRank (SSR) and SocialPageRank (SPR), which were designed to rank web pages by popularity. Experimentally they were able to confirm that user-created tags could sometimes improve web search, but the tags did not reliably provide sufficient information about a given web page. The variable contribution of user-created tags was attributed to tag ambiguity and, in a substantial number of cases, absence of user-created tags.

The FolkRank algorithm, devised by Hotho et al. [10], has become the most widely used method for ordering results in tag-based systems. They considered folksonomy systems that provided a search mechanism, but found that they lacked a formal approach to ranking, and simply displayed results by recency. Ranking methods used in search engines are not suitable for use with folksonomy systems as tags are typically very short and consist of just a few words. By contrast, major approaches to ranking require much more text from which to build indexes. Hotho et al. developed their FolkRank algorithm for folksonomies using an adaptation based on PageRank made famous by Google. FolkRank can be used to find communities in folksonomies by highlighting the top tags, resources and users, and by making this information explicit. Users with similar interests can get to know each other easily and share resources [10].

Using Hotho's algorithm 'FolkRank', Krause et al. [11] carried out an experiment as part of his research to compare URLs and ranking methods used in traditional web search engines and tag-based systems. FolkRank (Hotho et al. 2006) was compared to the TF-IDF and cosine similarity. Their experiment showed that FolkRank provides a

ranking order which is close to the ranking methods used in Google, and therefore gives the items most likely to be of relevance the greatest prominence [11].

## 4 Prototype Development

Based on Golder and Huberman [8], Michlmayr [14], Gay and Tonkin [9] and Hotho [10], in the literature review above we propose a method to improve tag-based search by enhancing the user-created tag set with automatically created words or phrases which describe the contents of the documents. Clearly, there is no standard method for automatically extracting the ‘best’ descriptive terms from a document, not least because the criteria for identifying the best terms may be very difficult to determine. However, our approach is a development of previous work in which we compared three alternative techniques for term extraction:

- Yahoo Term extraction API, which is a widely used but proprietary service.
- Selection of keywords and description meta tags.
- Selection of terms based on the highest word frequency, with extra weighting for words in the title of the document.

Experimental work showed that the Yahoo terms provide the greatest improvement in the search results [3], therefore in this paper we use the terms extracted from the original document using the Yahoo API. We will refer to the combination of user-created tags and tags provided by the Yahoo API as the enhanced tag set.

Our database consists of 73,000 records from the del.icio.us website. Del.icio.us [17], which started in 2003 and is one of the largest social bookmarking sites, is used by more than five million people and has 175 million bookmarked URLs tagged by different users [18]. For each document, a record holds the url, title, creator, date and the user-created tags. There were three major stages in developing the database. First, the html file was retrieved and saved, Second, an initial clean-up process was carried out on both the html files and user-created tags. Cleaning the html files involved removing unwanted characters and stop words and stripping out html tags whilst keeping the keywords and description meta tags to be used later on in the research. Cleaning the user-created tags included removing words containing numbers and non-English characters, and also stemming words using the Porter stemming algorithm. Third, for each record the top 5 keywords were automatically extracted using the Yahoo API term extraction tool. The decision to store the top 5 keywords for each document was taken after reviewing the full set of tags returned by the Yahoo system. Generally speaking, the keywords with the highest frequency conveyed the most document-specific information, so they should be the most effective keywords. However, as the frequency of keyword occurrence drops the ‘noise’ factor increases and the likelihood that such a keyword will represent document-specific information is reduced. The Yahoo terms for each document were then added to the database (see Table 1). Later, the ETS could combine these with the user-created tags to form the enhanced tag set.



**Table 1.** Example database record

Link	www.webteacher.com/javascript/index.html
Title	JavaScript tutorial for the total non-programmer
Tag-Set 1: User-created tags	Javascript, Good.site, learn
Tag-Set 2: Yahoo API	JavaScript tutorial, programming language, object oriented programming, non-programmer tutorial
Most Frequent	JavaScript, tutori, JavaScript tutori, program language

#### 4.1 Enhanced Tag Set Search Engine

The fundamental search algorithm in our search engine is based on co-occurrence between tags, and the results are ordered using the FolkRank method as the basis for selecting the query results. Table 2 briefly illustrates the concept of co-occurrence among records [1], and how related documents can be retrieved that do not explicitly contain any of the tags used as part of the query.

**Table 2.** Tag Co-occurrence example

1	Url-1	Australia	Sydney	Summer-Holiday
2	Url-2	Sydney Opera House	Australia	Tourist-places
3	Url-3	Harbour Bridge	Sydney Opera House	Beaches

In Table 2, the tag ‘Australia’ occurs in both record 1 and record 2, while ‘Sydney Opera House’ occurs in both record 2 and record 3. Because of the co-occurrence of ‘Australia’ and ‘Sydney Opera House’ in record 2, record 3 which is tagged with ‘Sydney Opera House’ would also be returned in a search using the tag ‘Australia’ even though record 3 is not tagged with ‘Australia’. The Enhanced Tag Set Search engine (ETS) was designed to allow a comparison to be made of searching our del.icio.us data set based on different tag sets, i.e. the user-created tags alone, and then those tags augmented by up to 5 terms from the Yahoo API service. The search algorithm selects urls that share similar tags based on tag co-occurrence.

The ETS search results were based on 2 different tag sets (user-created tag set and the enhanced tag set). The search algorithm was instrumented to employ, at any one time, either user-created tags only or the enhanced tag set, so that evaluation experiments could determine which tag set returned the set of results deemed most helpful by the participants.

In our algorithm we used a modified FolkRank approach to rank the search results. FolkRank deals with folksonomies as a “tuple  $F := (U, T, R, Y)$  where  $U$ ,  $T$ , and  $R$  are

finite sets, whose elements are called *users*, *tags* and *resources*, respectively, and  $Y$  is a ternary relation between them, whose elements are called tag assignments (TAS for short)" [10]. Users are identified by their IDs, the tags are strings, and the resources in our case are urls. This tuple can be presented in a undirected hypergraph where  $G = (V, E)$ ,  $V$  is the set of nodes and  $E$  is the set of hyperedges [10].

For the purpose of our study users are not of direct relevance as we concentrate on tags and how to enhance their value rather than finding communities of taggers with similar interests, therefore only the set of tags  $\{T\}$  and resources  $\{R\}$  were considered in our ranking approach. A weight is given to each resource and tag in our dataset to reflect how important they are as shown below:

$$Weight(r) = \frac{f(t)}{TF(r)} \quad (1)$$

$$Weight(t) = \frac{f(r)}{TF(t)} \quad (2)$$

In equation (1)  $r$  is the resource (url),  $f(t)$  refers to the frequency of resources tagged with the tag  $t$ ,  $TF(r)$  refers to the frequency with which the resources occur.  $Weight(r)$  is an estimate of how important a resource is in the database. It can take a value between 0 and 1, the greater the value the more important the resource.

In the equation (2),  $t$  is a user tag,  $f(r)$  refers to the frequency with which the tag was used to describe the resource  $r$ , and  $TF(t)$  refers to the frequency of the tag  $t$  in the whole database.  $Weight(t)$  is an estimate of how important a tag is relative to all other tags in the database. As with  $Weight(r)$  it can take a value between 0 and 1.

The algorithm allocates a high weight to an important resource tagged by an important tag. The importance of tags and resources is determined according to the number of co-occurrences for both the tags and the resources. If a tag is used many times in the database then a high weight is assigned to it and the same with the resources.

## 5 Evaluation

To compare the ETS's effectiveness when using user-created tags (Tag-Set1) with the enhanced tag set (Tag-Set2) we devised a lab-based experiment. The session began by giving the 42 participants an introduction to the experiment, outlining the aim and the search engine. A username and password for each participant was created so the results could be saved for each participant for later analysis. Participants were aware that the experiment consisted of two phases, but they were only told that the search engine was set up in a different way for each, so they were not biased to expect one to be superior. The order in which students were exposed to the two tag-sets was randomly allocated to cancel out any learning effect.

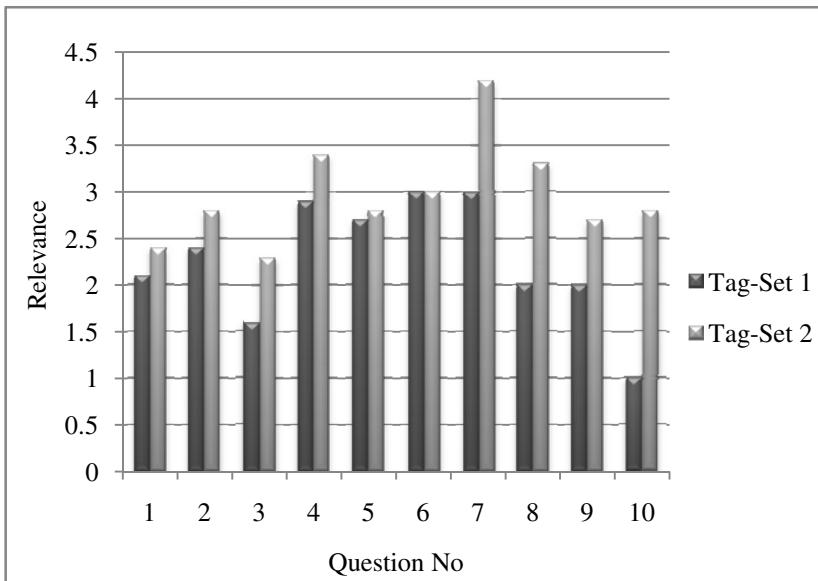
For both tag sets the participants were asked to use 10 different query phrases. These were supplied to them using our knowledge of the tags in the dataset. The participants used the same 10 search phrases in experimenting with both tag sets. This

was because free choice of query terms would likely have led to user frustration given the limited size of our dataset. Participants were asked to review reasonably carefully at least 10 of the search results for each query and to score, on a 1 – 5 scale, their impression of how relevant the results were to the initial query. At the end of the experiment a questionnaire was handed out to the participants to gather further information about users' views, recommendations and preferences.

## 6 Results

A statistical analysis was carried out to compare users' ratings of the search results using the user-created tags (Tag-Set 1) and the enhanced tag set (Tag-Set 2). Our hypothesis was that participants would find the enhanced tag set provided the more relevant search results. A detailed analysis of the relevance ratings reveals that users perceived real differences in the relevance of the results supplied by ETS using the 2 different tag-sets, and that invariably the enhanced tag sets outperformed, or at least did as well as, the user tags alone.

The average rating for the 10 queries using Tag-Set 2 is higher than the average rating for the queries using Tag-Set 1 except for query 6 which gives the same average rating for both phases (see figure 1). A paired t-test was used to compare the mean ratings for the two tag sets. As shown in table 3, with the exception of questions 5 and 6, the difference is statistically significant at the 5% level.



**Fig. 1.** Average Search Relevance

Table 4 shows paired t-test results for overall participant ratings, so that for each person, summing their scores over the 10 queries and comparing the totals, the mean difference was -8.39. This is significant at the 1% level and indicates that Tag-Set 2, the enhanced tag set, lead to search results that participants deemed superior.

**Table 3.** Paired t-test results for each question

Question	n	Tag-Set 1 Mean Rating	Tag-Set 2 Mean Rating	T	df	1-tailed p	Std. Error Diff.	Mean difference
1	42	2.0	2.4	-2.80	41	0.0039	0.14	-0.4
2	42	2.3	2.8	-4.16	41	0.0001	0.11	-0.5
3	42	1.5	2.3	-5.30	41	0.0001	0.14	-0.8
4	42	2.8	3.4	-4.66	41	0.0001	0.13	-0.6
5	42	2.6	2.8	-1.19	41	0.1206	0.14	-0.2
6	42	2.9	2.9	-0.10	41	0.4615	0.24	0
7	42	3.0	4.1	-5.44	41	0.0001	0.20	-1.1
8	42	2.3	3.2	-4.78	41	0.0001	0.18	-0.9
9	42	2.0	2.5	-3.14	41	0.0016	0.17	-0.5
10	42	1.1	2.6	-8.14	41	0.0001	0.18	-1.5

**Table 4.** Paired t-test results for overall participant ratings

N	Tag-Set 1 Mean Rating	Tag-Set 2 Mean Rating	T	df	1-tailed p	Std. Error Diff.	Mean difference
42	22.7	29	-8.39	41	0.0001	0.76	-6.4

## 7 Summary

In this paper, we have argued that user-created tags do not adequately or reliably describe document content, therefore there is a need to augment the user-created tags by adding more content-based terms.

We presented a method to enhance the user-created tags by adding metadata which is automatically extracted from the original document using the Yahoo API. We carried out an experiment comparing, 1) search based only on user-created tags with, 2) search using the enhanced tag set, and reported evidence that incorporating the extra tags can offer significant benefits. Experimental results showed that the relevance of search results based on user-created tags could be significantly improved when the set of tags was enhanced with content-based keywords.

To further benchmark ETS, a comparison of the search results could be performed between ETS and the search in another tag-based system, e.g. Delicious or StumbleUpon [20].

## References

1. Alag, S.: *Collective Intelligence in Action*. Manning, Greenwich CT (2008)
2. Al-Khalifa, H., Davis, H.: Folksonomies versus Automatic Keyword Extraction: An Empirical Study. In: *IADIS Web Applications and Research* (2006), <http://eprints.ecs.soton.ac.uk/14292/>
3. Awawdeh, R., Anderson, T.: Improved search in tag-Based systems. In: *ISDA 2009 - 9TH International Conference on Intelligent Systems Design and Applications*, Pisa, pp. 288–293 (2009)
4. Bao, S., Wu, X., Fei, B., Xu, G., Su, Z., Yu, Y.: Optimizing web search using social annotations. In: *Proc. of WWW 2007*, pp. 501–510 (2007)
5. Begelman, G.: *Automated Tag Clustering: Improving search and exploration in the tag space* (2006), [http://www.pui.ch/phred/automated\\_tag\\_clustering/automated\\_tag\\_clustering.pdf](http://www.pui.ch/phred/automated_tag_clustering/automated_tag_clustering.pdf)
6. Berendt, B., Hanser, C.: Tags are not Metadata, - but Just More Content – to Some People (2007), <http://www.icwsm.org/papers/2-Berendt-Hanser.pdf>
7. Brooks, C., Montanez, N.: Improved annotation of the blogspere via autotagging and hierarchical clustering. In: *WWW 2006: Proceedings of the 15th International Conference on World Wide Web*, Edinburgh, pp. 625–632 (2006)
8. Golder, S., Huberman, B.: The structure of collaborative tagging systems. HP labs (2006), <http://www.citeulike.org/user/zelig/article/305755>
9. Guy, M., Tonkin, E.: Folksonomies: Tidying up Tags. *D-Lib Magazine* (2006), <http://www.dlib.org/dlib/january06/guy/01guy.html>
10. Hotho, A., Jäschke, R., Schmitz, C., Stumme, G.: Information Retrieval in Folksonomies: Search and Ranking. In: Sure, Y., Domingue, J. (eds.) *ESWC 2006*. LNCS, vol. 4011, pp. 411–426. Springer, Heidelberg (2006)
11. Krause, B., Hotho, A., Stumme, G.: A comparison of Social Bookmarking with Traditional Search (2008)
12. Kipp, M.E.I.: Searching with Tags: Do Tags Help Users Find Things? In: *10th International Conference of the International Society for Knowledge Organization*, Montreal, Quebec, Canada (2008)
13. Macgregor, G., McCulloch, E.: Collaborative tagging as a knowledge organisation and resource discovery tool. *Library Review* 55(5), 291–300 (2006)
14. Michlmayr, E.: A Case Study on Emergent Semantics in Communities. In: Horrocks, I., Hendler, J. (eds.) *ISWC 2002*. LNCS, vol. 2342, Springer, Heidelberg (2002)
15. Quintarelli, E.: Folksonomies: power to the people (2005), <http://www.iskoi.org/doc/folksonomies.htm>
16. Yang, K.: Information retrieval on the Web. *Annual Review of Information Science and Technology* 39(1), 33–80 (2005)
17. <http://delicious.com/>
18. <http://www.ebizmba.com>
19. [http://developer.yahoo.net/blog/archives/2009/08/term\\_extraction\\_stays.html](http://developer.yahoo.net/blog/archives/2009/08/term_extraction_stays.html)
20. <http://www.stumbleupon.com/>

# Towards a Framework for Trusting the Automated Learning of Social Ontologies

Konstantinos Kotis<sup>1</sup>, Panos Alexopoulos<sup>2</sup>, and Andreas Papasalouros<sup>1,3</sup>

<sup>1</sup> University of the Aegean, Dept. of Information and Communication Systems Eng.,  
Ai-Lab, Samos, Greece 83200

kotis@aegean.gr

<sup>2</sup> IMC Technologies, Athens, Greece

palexopoulos@imc.com.gr

<sup>3</sup> University of the Aegean, Dept. of Mathematics, Samos, Greece 83200

andpapas@aegean.gr

**Abstract.** Automatically learned social ontologies are products of *social fermentation* between users that belong in communities of common interests (CoI), in open, collaborative and communicative environments. In such a setting, *social fermentation* ensures automatic encapsulation of agreement and trust of the shared knowledge of participating stakeholders during an ontology learning process. The paper discusses key issues for trusting the automated learning of social ontologies from social data and furthermore it presents a framework that aims to capture the interlinking of agreement, trust and the learned domain conceptualizations that are extracted from such a type of data. The motivation behind this work is an effort towards supporting the design of new methods for learning *trusted* ontologies from social content i.e. methods that aim to learn not only the domain conceptualizations but also the degree that agents (software and human) may trust them or not.

## 1 Introduction

Web, Social Web and even Semantic Web content can be reused for the creation of semantic content, shaping information into ontologies. However a critical mass of useful semantic content is missing. Web users can only find few well-maintained and up-to-date domain ontologies and the amount of RDF data publicly available is limited compared to the size of the unstructured Web information. Only a small number of Web users, typically members of the Semantic Web community, build and publish ontologies. To assist and motivate humans in becoming part of the Semantic Web movement and contribute their knowledge and time to create or refine/enrich useful ontologies there is need to boost semantic content creation by providing Web users with a “starting point of assistance” i.e. automatically learned ontologies.

Traditionally, the learning of ontologies involves the identification of domain-specific conceptualizations that are extracted from text documents or other semi-structured information sources e.g. lexicons, thesauruses. Such learned ontologies do not utilize any available social data that may be related to the domain-specific data e.g. ownership details (contributor, annotator or end-user), tags or argumentation/

dialogue items that have been used to comment, organize or disambiguate domain-specific information, querying information related to user clicks on retrieved information. Recently, the learning of ontologies has also involved social content that is mainly generated within Web 2.0 applications. Social content refers to various kinds of media content, publicly available, that are produced by Web users in a collaborative and communicative manner. Such content is associated to some social data that have been produced as a result of *social fermentation*. The most popular social data in Web 2.0 content is tags, which are (often) single words listed alphabetically and with a different font size or color (to capture its importance). Tags are usually hyperlinks that lead to a collection of items that are associated with. Such social data can be processed in an intelligent way towards shaping social content into ontologies. Since social data is produced as part of the social fermentation (tags are introduced in a collaborative and communicative manner), it can be argued that the learned ontologies that are produced from such a process encapsulate some degree of agreement and trust of the learned conceptualizations.

Social content generation (SCG) refers to a conversational, distributed mode of content generation, dissemination, and communication among communities of common interest (CoI). Social intelligence (SI) aims to derive actionable information from social content in context-rich application settings and to provide solution frameworks for applications that can benefit from the "wisdom of crowds" through the Web. Within this setting, a social ontology can be defined as: *an explicit, formal and commonly agreed representation of knowledge that is derived from both domain-specific and social data*. In the context of this chapter, the meaning of the term "social ontology" must be clearly distinguished from the meaning that is used in social sciences. A representative social-science definition is given by T. Lawson of the Cambridge Social Ontology Group<sup>1</sup>: "...the study of what is, or what exists, in the social domain; the study of social entities or social things; and the study of what all the social entities or things that are have in common".

Formally, an ontology is considered to be a pair  $O=(S, A)$ , where  $S$  is the ontological signature describing the vocabulary (i.e. the terms that lexicalize concepts and relations between concepts) and  $A$  is a set of ontological axioms, restricting the intended interpretations of the terms included in the signature [3], [4]. In other words,  $A$  includes the formal definitions of concepts and relations that are lexicalized by natural language terms in  $S$ . In this paper, we extend such model by a social dimension (equal to *social semantics*) that is influenced by the definition of "Actor-Concept-Instance model of ontologies" [7] formulated as a generic abstract model of semantic-social networks. The extended model is build on an implicit realization of emergent semantics, i.e. meaning must be depended on a community of agents. According to the extended model, a social ontology can be considered a triple  $O=(C, S, A)$ , where  $C$  is the set of collaborating contributors that have participated in a SCG task, from which  $S$  and  $A$  have been derived using the SI found in  $C$ . The range however of  $C$  over both  $S$  and  $A$  at the same time is not guaranteed, i.e.  $S$  may have been derived from  $C$ , but not  $A$ , which may have been automatically derived from external information sources such as a general ontology or lexicon e.g. from WordNet.

---

<sup>1</sup> T. Lawson, A Conception of Ontology, The Cambridge Social Ontology Group, 2004, [http://www.csog.group.cam.ac.uk/A\\_Conception\\_of\\_Ontology.pdf](http://www.csog.group.cam.ac.uk/A_Conception_of_Ontology.pdf)

The automated learning of social ontologies can be seen as a two-dimensional problem. The first dimension concerns the automated creation of ontologies from content (social and domain-specific), and the second, the social dimension, concerns collaboration and communication aspects (the *social fermentation*) that are involved during the creation of the content. Since automation is also involved, and human agents do not participate in the conceptualizations' agreement process, a key issue here is the trust on the extracted ontological agreement from social data i.e. the certainty that contributors of shared conceptualizations about a specific domain have agreed on a common understanding about the domain and that such agreement is successfully extracted in an automated fashion from social data (e.g. in open Web agents' world where agents must trust each others conceptualizations about the domain of discourse in order to be able to collaborate within an agreed context). In terms of the "trust the content" problem, the paper follows the assumption that the content used as input in an ontology learning process is a social one (or content that is involved in social fermentation), thus it is, at least in some degree, agreed and trusted. Blogs, (Semantic) Wikis, Folksonomies and other more sophisticated Web 2.0 applications such as Yahoo!Answers or Fixya.com, provide reputation-based trust (use personal experience or the experiences of others, possibly combined, to make a trust decision about an entity) or voting mechanisms for their content. Other types of content such as Web users' query logs provide a kind of trusting their content, based on the "majority vote of user clicks" on Web search results.

To the best of our knowledge and from literature review [1], [8], [10], currently there is no mean to automatically discover and attach uncertainty values on automatically learned social ontologies' signature ( $S$ ), axioms ( $A$ ) and contributors ( $C$ ). This paper proposes a model that represents trust for an ontology of the form  $O = \{C, S, A\}$ . More specifically, trust is formed as a meta-ontology which represents meta-information related to each element of a social ontology i.e. classes, properties, instances, contributors. Such meta-information is related to social data (e.g. contributors details, voting information) that is in turn interlinked to the content represented in the domain ontology. The definition of  $O = \{C, S, A\}$  is then extended, as shown in the paper, by introducing also trust representation.

The paper is structured as follows: section 2 presents the proposed framework for trusting social ontologies automatically learned by social content, section 3 reports on case studies for applying the proposed framework, and section 4 concludes the paper.

## 2 The Proposed Framework

### 2.1 Representing Trust in Social Ontologies

This paper proposes a model that represents trust for an ontology of the form  $O = \{C, S, A\}$ . More specifically, it is formed as a meta-ontology which represents meta-information related to each element of a social ontology i.e. classes, properties, instances, contributors. Such meta-information is related to social data (e.g. contributors details, voting information) that is in turn somehow related to the content represented in the domain ontology. The definition of  $O = \{C, S, A\}$  is then extended by introducing trust  $T$  for  $C$ ,  $S$  and  $A$  such as  $T = \{u, v_a, v_j\}$  where:  $u$  specifies the uncertainty



value computed for a given instance of  $C$ ,  $S$  or  $A$ ,  $v_a$  specifies the number of votes that do not trust an instance of  $C$ ,  $S$  or  $A$ , and  $v_f$  specifies the number of votes that do trust an instance of  $C$ ,  $S$  or  $A$ . In other words, some trusted (with some degree of uncertainty) contributors  $C$  are trusting (with some degree of uncertainty) a particular class, property or instance (i.e. an instance of  $S$  ontological signature) or an axiom (i.e. an instance of  $A$  axioms) that is learned from  $C$ 's contributed content. Although the computation of  $u$  (*uncertainty value*) reflects the trust in  $C$ ,  $S$  or  $A$  within a social network of  $C$  contributors,  $v_a$  and  $v_f$  values are reflecting the absolute *number of agreement* among the members of  $C$  for a given member of  $C$ ,  $S$  or  $A$ .

## 2.2 Integrate Trust in HCOME-3O Meta-ontologies Framework

Ontologies are *evolving* and *shared* artefacts that are collaboratively and iteratively developed, evolved, evaluated and discussed within communities of common interest (CoI), shaping domain-specific information spaces. To enhance the potential of information spaces to be collaboratively engineered and shaped into ontologies within and between different communities, these artefacts must be escorted with *all* the necessary meta-information concerning the conceptualization they realize, implementation decisions and their evolution. In HCOME-3O framework [11], the integration of three (meta-)ontologies that provide information concerning the conceptualization and the development of domain ontologies, the atomic changes made by knowledge workers, the long-term evolutions and argumentations behind decisions taken during the lifecycle of an ontology, has been proposed (and evaluated via its utilization in later work). This involves ontology engineering tasks for a *domain* ontology and its versions (*domain knowledge*), i.e. editing, argumentation, exploiting and inspecting, during which meta-information is captured and recorded (*development ontologies*) either as information concerning a simple task or as information concerning the inter-linking of tasks. This framework has been proposed in the context of HCOME collaborative engineering methodology [5].

Recently, HCOME methodology has been extended with ontology learning tasks [5] in order to capture knowledge that is automatically extracted from content and learned in the domain ontology. In such a new dimension of the methodological aspect of ontology engineering, agent agreement on automatically learned conceptualizations may be assisted by integrating representations of already computed uncertainty values in the following way: collaborating knowledge contributors consult uncertainty values of the learned conceptualizations and agree or disagree on the conceptualizations.

The integration of the proposed model into the HCOME-3O framework can be easily achieved by merging its semantics with the Administration meta-ontology [11], which mainly records instances of domain conceptualizations (classes, properties, individuals) and contributors of such conceptualizations, in the following way (Figure 1): a) add trust-related datatype properties (“*uncertainty\_value*”, “*votes\_against*”, “*votes\_for*”) of the trust model to the Administration meta-ontology, Administered\_Item class), b) add object properties (*has\_superClass*, *has\_Domain*, *has\_Range*, *has\_Type*) to the corresponded ontology elements, extending the Administration meta-ontology, in order to facilitate the assignment of trust on (simple) axioms ( $A$ ) also.

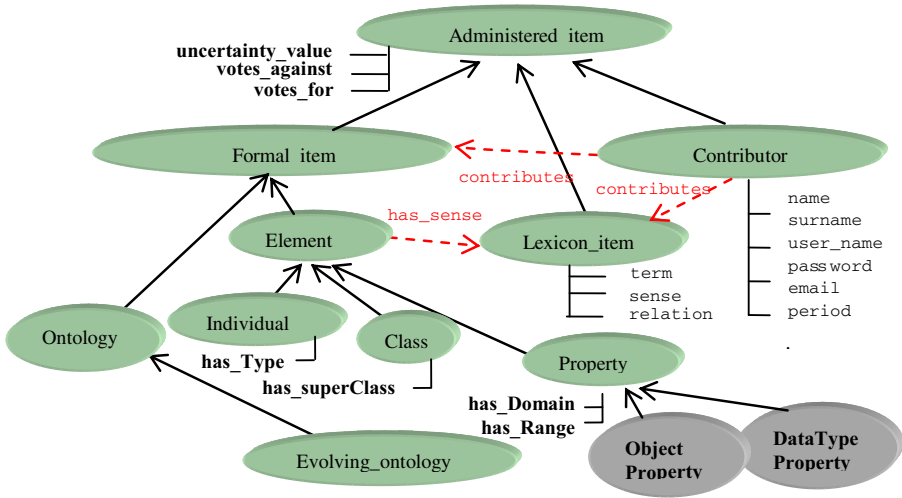


Fig. 1. The proposed trust meta-ontology integrated in HCOME-30

A *Formal Item* (a Class, a Property or an Individual item of the domain ontology) is recorded in the meta-ontology as domain knowledge that is contributed by a specific Contributor. Trust-related properties are attached to formal items, represented with the dataType properties: *uncertainty\_value*, *votes\_against*, *votes\_for*. Such properties are inherited to all formal items of the extracted signature  $S$  of the domain ontology. Furthermore, the trust-related properties are also attached (inherited via the Administered Item specification) to the Contributor. Such information is necessary in order to keep record of trusted (by others) people also. A similar conceptualization for trusting people is provided in Trust ontology of MindSwap (<http://trust.mindswap.org/trustOnt.shtml>) however it is not interlinked with trust-aware domain conceptualizations that these people may contribute. On the other hand, several other efforts have been lately presented (see related work section) for modeling trusted conceptualizations; however they are not interlinked with the trust-aware recording of their contributors. Having said that, since the Administration meta-ontology is part of a wider meta-information framework (that is HCOME-30 framework), the ‘range’ of the trust-related properties can be expanded to other meta-information also such as the changes that are recorded between each new version that contributors are developing, the argumentation dialogue items (arguments, issues, positions) that are recording during the collaborative evaluation/development of ontologies, etc. As a result of this effect, the model can provide answers to more complex and trust-elaborated queries such as ‘give me all changes that were made between version  $O_i$  and version  $O_{i+1}$  of the ontology  $O$ , by contributors with a trust  $T = \{0.6, 4, 5\}$ ’ or ‘give me all suggesting positions (argumentation items) that were made by the community for classes of the domain ontology with a trust  $T = \{0.6, 4, 5\}$ ’.

For a detailed description of the Administration meta-ontology and its role in the HCOME-30 framework please refer to the related article [11].

Trust on ontological axioms ( $A$ ) can be also extracted since they also comprise domain knowledge that can be discovered in social content. Trust on axioms however can be easily inferred from already trusted classes (upper level) and simple axioms (e.g. subsumption) of the learned ontology. Consider for instance axioms of a simple pre-specified topics' hierarchy (e.g. the one in Yahoo!Answer Web 2.0 application): If class  $A$ , class  $B$  and class  $C$  are trusted with value "1.0" and the axioms  $A \sqsubseteq B$  and  $B \sqsubseteq C$  are also trusted with values "1.0", then the inferred axiom  $A \sqsubseteq C$  can be also trusted with value "1".

Concluding the paragraph, the definition of the social ontology can now be reformulated as follows:  $O = \{T, C, S, A\}$  where  $T = \{u, v_a, v_p\}$  is the trust function  $T : C \cup S \cup A \rightarrow [0,1] \times Z \times Z$  for ontological signature  $S$  and axioms  $A$  that a set of collaborated contributors  $C$  participated in a task of social content generation (SCG) have derived based on their social intelligence (SI).

### 2.3 Computation of Uncertainty Values

For a Web social application that uses a voting system to trust (or not) some content (e.g. in Yahoo!Answers application where users vote for or against a posted answer to a Yahoo!Answers community question), the uncertainty value  $u$  for this chunk of content can be computed using a simple formula  $u = v_f - v_a$ , i.e. the number of votes after subtracting the votes against ( $v_a$ ) from the votes for ( $v_f$ ). The vector of the voting values computed for some content,  $U = (u_1, u_2 \dots u_n)$  where  $n$  represents the number of content chunks that have been related to the voting system (e.g. in the Yahoo!Answers application  $n$  is the number of answers  $a$  posted for a question  $q$ ), is then normalized to the interval  $[0, 1]$ . The normalization will work well if data is positive or zero. If data contains negative numbers, for example,  $-1, 3$  and  $4$ , then the sum is 6. If it is normalized by the maximum value we get  $-1/6, 1/2$ , and  $2/3$ . The sum of the three values is still 1 but now a negative number ( $-1/6$ ) is part of the index. The following general solution may be however applied: Shift data by adding all numbers with the absolute of the most negative (minimum value of data) such that the most negative one will become zero and all other numbers become positive. Then data is normalized using any common normalization method for zero or positive numbers. For example, if data is  $-1, 3$  and  $4$ , the most negative number is  $-1$ , thus we add all numbers with  $+1$  to become:  $0, 4, 5$  and then normalize it.

A general normalization solution for voting values in a social application is proposed in the following lines. Suppose we have a range or scale from  $A$  to  $B$  and we want to convert it to a scale of 1 to 10, where  $A$  maps to 1 and  $B$  maps to 10. Furthermore, we want to do this with a linear function, so that for example the point midway between  $A$  and  $B$  maps to halfway between 1 and 10, or 5.5. Then the following (linear) equation can be applied to any number  $x$  on the  $A$ - $B$  scale:

$$y = 1 + (x - A) * (10 - 1) / (B - A). \quad (1)$$

Note that if  $x = A$ , this gives  $y = 1 + 0 = 1$  as required, and if  $x = B$ , then:

$$y = 1 + (B - A) * (10 - 1) / (B - A) = 1 + 10 - 1 = 10, \quad (2)$$

as required. One can use this equation even if  $A > B$ . In our case, the scale will be 0.0 to 1.0 for every  $x$ , where  $x \in \{u\}$ .

## 2.4 Using the Framework for Automatically Generating Fuzzy Ontologies

The fact that the trust model of the presented framework assigns uncertainty values to the elements of the ontology learned through the *social fermentation* process practically means that (most of) the knowledge represented by this ontology is uncertain. Typically, representation of uncertain knowledge is facilitated by *fuzzy ontologies*, namely ontologies that utilize the notions of *fuzzy set* and *fuzzy relation* [6] in order to suggest that certain pieces of knowledge should be considered as true at certain degrees.

As with traditional ontologies, the pure manual generation of a fuzzy ontology is a difficult and tedious task that requires the active involvement of domain experts, mainly for the task of assigning truth degrees to the ontology's elements. Since our framework provides a way for automatically performing this task, we claim that it may be as well used for the automatic generation of fuzzy ontologies. To show why this is the case we consider a formal definition of a fuzzy ontology, adapted from [12], in which the latter is a tuple  $O_F = \{C, I, FR, FA, FLV, FVA\}$  where:

- $C$  is a set of concepts (classes) and  $I$  is a set of individuals.
- $FR$  is a set of fuzzy relations. Each fuzzy relation is a function  $E^2 \rightarrow [0,1]$  where  $E$  is the union of  $C$  and  $I$ . Of particular importance are two fuzzy relations: the fuzzy subsumption relation between concepts and the fuzzy instantiation relation between concepts and instances.
- $FA$  is a set of fuzzy attributes. Each fuzzy attribute is a function  $I \rightarrow F(X)$ ,  $F(X)$  being the set of all fuzzy sets in the universe of discourse  $X$ .
- $FLV$  is a set of fuzzy linguistic variables. Each variable is a tuple  $\{u, T, X, m\}$  in which  $u$  is the name of the variable,  $T$  is the set of linguistic terms of  $u$  that refer to a base variable whose values range over a universal set  $X$  and  $m$  is a semantic rule that assigns to each linguistic term a meaning in the form of a fuzzy set in  $X$ .
- $FVA$  is a set of fuzzy valued attributes. Each fuzzy valued attribute is a function  $I \rightarrow T$  where  $T$  is the set of the linguistic terms of a fuzzy linguistic variable.

Given this definition, generating a fuzzy ontology practically means assigning truth degrees to fuzzy relations and defining the meanings of fuzzy linguistic terms. As can be seen from figure 1, our framework supports the first from these two tasks through the assignment of uncertainty values to the *has\_superclass* property of the *Class* item (fuzzy subsumption), the *has\_type* property of the *Individual* item (fuzzy instantiation) and to the instances of the *Object* and *Datatype* property items (fuzzy relations and fuzzy attributes). The support of the second task, namely generation of linguistic term meanings, is left as future work as it requires an extension of the administration meta-ontology to the fuzzy realm.

Related work on the learning of fuzzy ontologies comprises methods that perform text mining in order to generate degrees for the fuzzy subsumption relation between concepts [6] [9] and for the fuzzy instantiation relation [6]. The work presented in this

paper differs from these approaches in two ways. First of all it is more complete as it supports the automatic generation of any fuzzy relation, not only of the fuzzy subsumption and instantiation ones, as well as of fuzzy attributes. In addition to that, however, it follows a different perspective as uncertainty in the learned ontologies is not captured as an effect of a “good” or “bad” text mining technique but it is rather a result of the social fermentation process during the creation of social content (social data and domain-specific content). This does not only seem to be the right approach when referring to social ontology learning but it is in-line with the social dimension of the automatic learning ontology process.

Of course, the text-mining uncertainty dimension may be also of some importance when combined with a social one: A “gold” approach towards trusting automated learning of social ontologies can serve as a merger of both dimensions i.e. the text-mining and the social one. Intuitively, the average uncertainty of the two values can be considered the “gold” uncertainty value  $u_g$  of the formulae  $T = \{u, v_a, v_f\}$  of our approach. However, more sophisticated formulas may be proposed, if based, for instance, on the work of learning trust decision strategies in agent-based reputation exchange networks [2], [8]. Assuming that a pessimistic strategy is followed [8], where agents do not trust each other unless there is a reason to do so, the uncertainty value of a text mining approach should be weighted more than the value of a social one.

### 3 Case Studies

In order to evaluate the proposed framework, it is necessary to develop and use ontology learning methods that learn social ontologies as a result of a *social fermentation* process. For this purpose we have re-used the in-house recently developed ontology learning method which utilizes (for input) mined domain-specific query logs of Web users community [5] and we are in the process of implementing an additional ontology learning method that utilizes Web 2.0 social content from Web Question/Answers applications such as Yahoo!Answers.

To apply the proposed trust framework on the Queries-to-Ontology learning method, an important assumption has been made since in such a context a voting mechanism is not present. The formula of  $T = \{u, v_a, v_f\}$  is reduced to  $T = \{u\}$  since in this case  $v_a$  and  $v_f$  can be considered of zero value. The computation of  $u$  for a Web query  $q$  is based on the reputation of the query in a particular context. Such reputation is reflected by the number of clicks  $D\_click(q)$  on resulted documents  $D$  for a particular query  $q$  (reflecting that users’ interests have been found in this query). Since this value can be considered as the reputation of a particular query, it can also be considered as the reputation of the learned conceptualizations from the particular query that a contributor  $C$  provided, i.e. the query-related signature ( $S$ ) and axioms ( $A$ ) of the learned ontology. Thus, the formula  $O = \{T, C, S, A\}$  is valid for this use case. Low  $T$  values will be returned for low  $D\_click(q)$  values i.e. many Web users did not find search results to be much related to the query (they did not clicked on them). An additional step to this approach may be the analysis of history of queries: measuring the frequency of similar queries placed for the same context. This is left for future research.

Extending the work conducted using query logs as input to a social ontology learning process, a future direction is proposed in this paper, with the aim to trust the learning of social ontologies from Web 2.0 content. As a case study it was decided to apply the proposed framework on social content that is created by Yahoo! Answers community (an alternative is Fixya.com). Yahoo! Answers (<http://answers.yahoo.com/>) is a shared place where people collaborate and communicate by asking and answering questions on any topic. The aim of such a social fermentation is to build an open and commonly agreed knowledge base for the benefit of the community. Organized in topics (simple thematic category hierarchy), questions are posted by the users of the social network, expecting several answers that will eventually satisfy their knowledge acquisition needs. A voting for the best answer mechanism ensures that an agreed (by the majority) and trusted (by the number of “for” voters) answer is related to a question. Professional knowledge can also be shared within the community by *knowledge partners*. Such knowledge supplements the answers received from the community by answering questions in a specialized field, drawing on partners training, their professional experiences, and other appropriate resources. As a benefit, knowledge partners may mention their products or services, where relevant, in an answer, for advertisement reasons. Such a mutual benefit (for partners and community users) can guarantee a live social network that is difficult to “die” and at the same time it can guarantee the strong building of trust for the content that both stakeholders are sharing. The proposed method utilized the following inputs:

- 1) A question/answer document which contains the following information:
  - a. the topic of the question (and the more general/specific categories of the topic hierarchy). Topics are pre-defined by Yahoo!Answers application
  - b. user information: who posted the question, who posted an answer, who voted against or for
  - c. the question and the associated answers in natural language: users can post a title and a comment for the question, and only comments for their answers
  - d. the best answer and the votes for
  - e. the votes for all other answers
  - f. other related questions, resolved or open, on the same topic
- 2) WordNet lexicon. It will be used to enrich the ontology with additional semantics (entities, semantic relations, individuals)

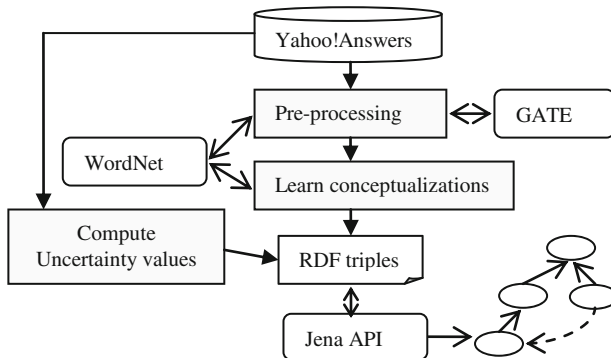
The processing of the proposed social ontology learning method, integrated with the proposed trust framework, is outlined in the following steps (Figure 2):

- *Step-1*: The method learns the starting RDF triples from the types of the pre-defined hierarchy that the topic of the posted question is classified under.
- *Step-2*: The posted question (both title and comment) is analyzed using an NLP API (e.g. GATE<sup>2</sup>) in order to identify parts of speech (POS) and perform tokenization.

---

<sup>2</sup> <http://gate.ac.uk/>

- *Step-3*: Since context is known (from Step-1) and some text analysis has been done (in Step-2), important terms can be identified and semantic relations between them can be recognized [5]. The following techniques can be used in combination:
  - a. Hearst patterns
  - b. Simple heuristic rules that utilize knowledge from the POS tagging.
- *Step-4*: Semantics are enriched using WordNet. Mapping of terms to WordNet senses is performed automatically using a statistical technique from Information Retrieval to compute latency of terms in term-document spaces (LSI method [6]).
- *Step-5*: Steps from Step-2 to Step-4 are repeated for the best (voted) posted answer. The ontology elements extracted from this step (classes, properties, instances) are assigned the uncertainty value 1.0 (representing the uncertainty of this element in respect to the community trust of the commonly agreed “best answer”).
- *Step-6*: Steps from Step-2 to Step-4 are repeated for the rest posted answers. To keep the size of the learned ontology low (and to avoid noise) only important terms (most frequent terms) are introduced as classes of the learned ontology. The importance of terms is a threshold value that can be empirically set at ‘2’. However, in large sized answers (more than one paragraph of text) such value must be set higher. Other techniques should be also tested to avoid noise of large answers (e.g. to first locate important partitions of the text, applying n-grams analysis for instance, and then extract important terms from there). The ontology elements extracted from this step (classes, properties, instances) are assigned an uncertainty value (normalized) between the interval 0 and 0.9.
- *Step-7*: The generated RDF triples from steps Step-2 to Step-6 are transformed into a consistent OWL model. The development proposed is based on Jena API and Pellet.



**Fig. 2.** The architecture of the proposed learning method

The output of the method is a learned ontology with uncertainty weights attached to its elements (classes, properties, instances). To respect the formula  $O = \{T, C, S, A\}$ , the learned ontology is recorded in the extended Administration meta-ontology of the HCOME-3O model, interlinking the trusted conceptualizations with trusted contributors. In this use case, the contributors are Yahoo!Answers voters, members of the Yahoo! community, for which trust values can be also computed using a) a point system that is provided by the application in order to represent the reputation in the community, and b) their experience in the community (time of registration).

The voting mechanisms integrated in Yahoo!Answers as well as in other Web 2.0 related applications (e.g. Fixya.com) provide social data that is able to relate some content i.e. a posted answer, to some other content, i.e. to a posted question, and to their contributors. Such interlinking can be interpreted as agreement or disagreement on users' opinion and eventually as a trust value of the shared knowledge that is encapsulated in the most agreed opinion (best voted answer). Trusted more or less, the related-to-a-topic knowledge is shaped into a domain ontology where each element is eventually associated with an uncertainty value that is computed directly from the social data associated with the represented content. Professional knowledge can also be shared within the community by *knowledge partners*. Such knowledge supplements the answers received from the community. Since this kind of knowledge is contributed by experts, it can be considered as highly trusted. Furthermore, the mutual benefit of knowledge partners and community users (advertisement and expertise knowledge contribution) plays a key role to "truth telling" when it comes to partners' answers in community users' posts. This can guarantee a live social network with strong roots of trust for the content that all stakeholders are sharing. Relatively to ontology learning from query logs method, the proposed ontology learning method can be trusted in a higher degree since its social data is both directly and indirectly associated with the content represented in the ontology.

## 4 Conclusions

This paper presents an effort towards devising a framework for trusting automatically learned social ontologies as part of a *social fermentation* between users that belong in communities of common interests (CoI), in open, collaborative and communicative environments. The paper discusses key issues towards this goal and focuses on the presentation of the model that interlinks agreement and trust with the learned domain conceptualizations that are extracted from social data of Web applications. The reported work contributes in the design of new ontology learning methods in a Web of trusted conceptualizations and their contributors. More specifically, the proposed framework can be used for consultation during the design of ontology learning from social data methods that need to automatically learn not only the domain conceptualizations but also the degree that agents trust these conceptualizations (and their contributors) or not.

## References

1. Artz, D., Gil, Y.: A survey of trust in computer science and the Semantic Web. *Journal of Web Semantics: Science, Services and Agents on the World Wide Web* 5, 58–71 (2007)
2. Fullam, K., Barber, S.: Learning Trust Strategies in Reputation Exchange Networks. In: AAMAS 2006, Hakodate, Hokkaido, Japan, May 8-12 (2007)



3. Kalfoglou, Y., Schorlemmer, M.: Ontology mapping: the state of the art. *The Knowledge Engineering Review* 18, 1–31 (2003)
4. Kotis, K., Vouros, G., Stergiou, K.: Towards Automatic Merging of Domain Ontologies: The HCONE-merge approach. *Journal of Web Semantics* 4, 60–79 (2006)
5. Kotis, K., Papasalouros, A.: Learning useful kick-off ontologies from Query Logs: HCOME revised. In: 4th International Conference on Complex, Intelligent and Software Intensive Systems (2010)
6. Lau, R., Li, Y., Xu, Y.: Mining Fuzzy Domain Ontology from Textual Databases. In: IEEE/WIC/ACM International Conference on Web Intelligence (2007)
7. Mika, P.: Ontologies are us: A unified model of social networks and semantics. *Journal of Web Semantics* 5, 5–15 (2007)
8. O’Hara, K., Alani, H., Kalfoglou, Y., Shadbolt, N.: Trust strategies for the semantic web. In: Proceedings of Workshop on Trust, Security, and Reputation on the SemanticWeb, 3rd International Semantic Web Conference (2004)
9. Tho, Q.T., Hui, S.C., Fong, A.C.M., Cao, T.H.: Automatic Fuzzy Ontology Generation for Semantic Web. *IEEE Transactions on Knowledge and Data Engineering* 18, 842–856 (2006)
10. Tang, J., Leung, H., Luo, Q., Chen, D., Gong, J.: Towards Ontology Learning from Folksonomies. In: JCAI 2009, Pasadena, California, USA, July 11-17 (2009)
11. Vouros, G., Kotis, K., Chalkiopoulos, C., Lelli, N.: The HCOME-3O Framework for Supporting the Collaborative Engineering of Evolving Ontologies. In: ESOE 2007 International Workshop on Emergent Semantics and Ontology Evolution, ISWC (2007)
12. Alexopoulos, P., Wallace, M., Kafentzis, K., Askounis, D.: Utilizing Imprecise Knowledge in Ontology-based CBR Systems through Fuzzy Algebra. *International Journal of Fuzzy Systems, Special Issue on Fuzzy Approaches for Ontology Applications and Adaptive Web Services* (in press)

# PlayPhysics: An Emotional Games Learning Environment for Teaching Physics

Karla Muñoz<sup>1</sup>, Paul Mc Kevitt<sup>1</sup>, Tom Lunney<sup>1</sup>, Julieta Noguez<sup>2</sup>, and Luis Neri<sup>2</sup>

<sup>1</sup> Intelligent Systems Research Centre, Faculty of Computing and Engineering, University of Ulster, Magee, BT48 7JL, Derry/Londonderry, Northern Ireland, UK

<sup>2</sup> School of Engineering and Architecture, Tecnológico de Monterrey, Mexico City, Col. Ejidos de Huipulco, Tlalpan, C.P. 14380, Mexico  
munoz\_esquivel-k@email.ulster.ac.uk,  
{p.mckevitt,tf.lunney}@ulster.ac.uk, {jnoguez,neri}@itesm.mx

**Abstract.** To ensure learning, game-based learning environments must incorporate assessment mechanisms, e.g. Intelligent Tutoring Systems (ITSs). ITSs are focused on recognising and influencing the learner's emotional or motivational states. This research focuses on designing and implementing an affective student model for intelligent gaming, which reasons about the learner's emotional state from cognitive and motivational variables using observable behaviour. A Probabilistic Relational Models (PRMs) approach is employed to derive Dynamic Bayesian Networks (DBNs). The model uses the Control-Value theory of 'achievement emotions' as a basis. A preliminary test was conducted to recognise the students' prospective-outcome emotions with results presented and discussed. *PlayPhysics* is an emotional games learning environment for teaching Physics. Once the affective student model proves effective it will be incorporated into *PlayPhysics*' architecture. The design, evaluation and post-evaluation of *PlayPhysics* are also discussed. Future work will focus on evaluating the affective student model with a larger population of students, and on providing affective feedback.

**Keywords:** Affective Student Modelling, Control-Value Theory, Dynamic Bayesian Networks (DBNs), Game-based Learning Environments, Intelligent Tutoring Systems, PlayPhysics, Probabilistic Relational Models (PRMs).

## 1 Introduction

Information Technology (IT) has influenced the world that surrounds young learners. They expect to feel motivated and engaged when learning [1]. Traditional education is evolving into a student centered-approach, which will support the development of learner's skills, e.g. creativity, self-organization and decision making [2]. Game-based learning environments are being used to actively involve students in their own learning [3]. Intelligent tutoring systems (ITSs) are assessment mechanisms that are included in the architectures of game-based learning environments to ensure students' understanding [4-5]. ITSs follow the learner's performance, identify the learner's needs and provide suitable feedback. The learners' emotional state has proven to be

deeply interrelated with their motivation and cognition [6]. Affective Computing and Affective Gaming seek to identify and influence the user's emotion [7-8]. Incorporating an affective dimension into ITSs involves the challenges of how to reason about and respond to the learner's emotion. Here we focus mainly on the former. The latter may involve multimodal output modulation.

An ITS is comprised of several modules. The student model enables an ITS to understand the learner's behaviour [9], whilst a tutor model selects the most suitable pedagogical action and the most effective way of conveying the teaching message [10]. Affective student modelling is a task that involves uncertainty, since diverse social, personal and cognitive factors influence the learners' emotional state [6]. Identifying whether students' emotions facilitate or impede learning will provide ITSs with the capability of ignoring or changing the learner's disposition to enhance understanding [7]. To date, there is no system that can identify all the relevant emotions that take place in a teaching and learning context. In addition, an affective student model that effectively reasons about the learner's emotional state using motivational and cognitive variables does not yet exist. Therefore, here we propose and discuss the design of such an affective student model, which uses the Control-Value theory of 'achievement emotions' as a basis, and uses observable behaviour and answers to posed questions as part of the game dialogue as evidence to update its knowledge. A Probabilistic Relational Models (PRMs) approach and Dynamic Bayesian Networks (DBNs) are employed for the design and implementation of the affective student model.

To evaluate the affective student model and the approach, a prototyping exercise will be carried out with undergraduate students enrolled in an Introductory Physics course. The prototyping exercise uses a "Wizard-of-Oz" experiment [11] as a basis. A preliminary test was carried out with students at postgraduate level to identify possible weaknesses in the prototyping material, and to evaluate the accuracy of the affective student model when reasoning about the students' outcome prospective emotions and results are discussed and presented. Once the effectiveness of the affective student model is ensured, it will be incorporated into *PlayPhysics'* architecture, *Olympia* [12]. The *Olympia* architecture will be modified to select the pedagogical actions, e.g. motivational, affective or cognitive, that maximize learning. It is important to signal that motivational, affective and cognitive strategies can be independent, complementary or in contraposition [13]. Therefore, selecting the most suitable pedagogical action or actions and finding the most suitable way of communicating errors to the students are also considered challenges. *Olympia* will provide pedagogical feedback through modulating game elements, e.g. game-characters, sounds and colours.

Students frequently experience problems when trying to understand the underlying principles of Physics [12]. *PlayPhysics* is an emotional game-based learning environment for teaching Physics at undergraduate level. Section 2 reviews state of the art research related to the challenges of reasoning about the learner's emotion and influencing the learner's emotional state. Section 3 describes the design and implementation of the affective student model. Section 4 describes *PlayPhysics'* design, implementation and future evaluation. Section 5 presents and discusses the results of the preliminary evaluation of the affective student model. Section 6 concludes by outlining the benefits of this research and describing its future work.

## 2 Background

To enhance learning and engagement, computer tutoring must be able to identify and influence the learner's emotional state. This section discusses the state of the art related to both challenges.

### 2.1 Identifying Emotion

ITSs are being updated to enable them to recognize the learner's motivation or emotion. Approaches identified in the field of ITSs are: (1) recognizing the physical effects of emotions [14], (2) predicting emotion from its origin [15] and (3) a hybrid approach derived from the previous two approaches [4]. Recognizing the physical effects involves using additional hardware to recognize gestures, body position, prosodic features and psycho-physiological data that is mapped to emotional meanings. Limitations of this approach are the acquiring of large quantities of data and also that hardware is intrusive and prone to failure. Predicting emotion from its origin is an approach that involves using a cognitive theory to reason about the possible causes of an emotion. Research is usually based on the Ortony, Clore and Collins (OCC) model, which classifies emotions according to their sources [16]. This theory must adapt to the learning context and account for the learner's attitudes, standards and beliefs. A hybrid approach predicts the existence of an emotion and matches data patterns to ensure that the emotion actually happened.

To recognize the motivational state of the learner using observable behaviour, an approach using quantitative and qualitative characteristics has proven effective [17]. The approach focuses on analyzing the learner's actions and mapping them to values associated with effort, independence and confidence. This work showed that self-efficacy could effectively infer the learner's disposition whilst learning. A self-efficacy model was derived from observable behaviour and physiological data [18]. Results showed that the model achieved 70% accuracy and this accuracy was increased by 10% by adding physiological data. To date, there is no ITS capable of identifying all students' emotions in a learning context. Therefore our research focuses on creating an affective student model using as a basis the 'Control-Value theory of Achievement Emotions'.

The Control-value theory of 'achievement emotions' is an integrative theory of emotion that considers control and value appraisals as the most relevant factors when determining an emotion [6]. 'Achievement emotions' are emotions that take place when academic or achievement activities are performed and a desired outcome is expected as a result of performing these activities. These emotions are domain dependent. *Control* can be defined as the perceived *control* over a specific activity and its outcomes, e.g. self-efficacy, before starting to solve a specific Physics or Mathematics problem. *Value* is the perceived importance of the outcome and the desirability of the activity, e.g. the learner's intention or interest. The Control-Value theory, instead of opposing the OCC model, incorporates some assumptions of that theory. The Control-Value theory classifies emotions according to their focus and time-frame: *outcome-prospective*, *activity* and *outcome-retrospective* emotions. For example, anticipatory joy, hope, anxiety, anticipatory relief and hopelessness are considered *outcome-prospective* emotions. The Control-Value Theory uses the Achievement Emotions Questionnaire (AEQ), a self-report tool designed and evaluated through

Structural Equation Modelling (SEM) [19]. The AEQ has been used to assess accurately the learners' achievement emotions in the English, German, Mathematics and Physics domains.

## 2.2 Influencing and Modulating Emotion

Systems are embodying experts, synthetic characters or embodied pedagogical agents (EPAs), in diverse fields with the objective of enhancing the effectiveness of the message [20]. The challenge is creating in the user a feeling of believability. To achieve this aim, research has focused on domain knowledge representation, emotional intelligence, personality, social relations and common sense.

Some ITSs use an EPA to communicate their pedagogical, motivational and affective responses [4-5, 14]. In these cases the response is accompanied by the inferred data of the student model. Herman, the EPA of the virtual learning environment (VLE) 'Design a Plant' [21], synchronises multiple modalities addressing context and time. Herman's behaviours are assembled in real time with segments of animations and audio using a sequencing-engine. To facilitate access to the behaviour space, the authors employed ontological, intentional and rhetorical indexes for handling explanatory, advisory and emotional behaviours. A hierarchy of pre-requisites and dependencies between behaviours was defined, as in The Oz Project [22]. To enable the EPA to move in the virtual environment, compositional and full-body animations may be employed, e.g. Cosmo's behaviours [23]. Some EPA architectures, e.g. Cosmo's architecture [23], are very similar to the architecture of an ITS. Research into Embodied Characters has created several visual languages, with their own syntax and semantics, to attain animation coherence. Examples of applications using a visual language are BEAT [24], Smartkom [25] and the Oz Project [22].

The control of the system over the character can be modelled as a central management entity that controls which actions are performed when and where, as is common with video games, or indicating the general direction of the interaction or narrative, but not the actions, as is common with interactive dramas [26]. Cognitive theories of emotion and personality theories have been used as a basis to create the reactive behaviour of embodied characters and EPAs. Theories employed commonly are the OCC model and the Big Five, as in The Oz Project [22], Fear Not! [27] and PrimeClimb [4].

Video games are multi-sensorial environments, where acoustic and visual sources serve diverse purposes [28-29], e.g. setting a mood, indicating changes in narrative, creating a feeling of immersion, focusing attention, conveying meaning, identifying objects and actors and decreasing the player's learning curve, therefore impacting on the reception of a video game. Colours have a diverse range of meanings, from the cultural to the personal as well as possessing innate meaning [30]. Colours have been employed for self-reporting of emotion in the classroom [31] and in virtual learning environments [32], and to communicate and emphasize the existence of an emotion [33].

## 3 Affective Student Model Design and Implementation

Student modelling is an undertaking that involves domains with inherent uncertainty [9], since it is not clear how the learner achieves knowledge and how personal

differences, social standards and the knowledge domain influence the learner's emotion and motivation. Selecting the data to be considered when updating the model also constitutes a challenge. Bayesian Networks (BNs) are employed to handle uncertainty and represent causal relations between random variables. They have been applied to implement cognitive, affective and motivational student models [12, 4-5].

**Table 1.** Summary of the Control-Value Theory by Pekrun et al. [6]

Time frame/focus on	Value appraisal	Control appraisal	Emotion
Prospective/outcome	Positive Success	High	Anticipatory Joy
		Medium	Hope
		Low	Hopelessness
	Negative Failure	Low	Hopelessness
		Medium	Anxiety
Retrospective/outcome	Positive Success	High	Anticipatory relief
		Irrelevant	Joy
		Self	Pride
		Other	Gratitude
	Negative Failure	Other	Anger
		Self	Shame
		Irrelevant	Sadness
Present/activity	Positive	High	Enjoyment
	Positive/Negative	Low	Frustration
	None	High/Low	Boredom
	Negative	High	Anger

To facilitate BN design, Probabilistic Relational Models (PRMs), object-oriented representations of the knowledge domain, can be employed to overcome the limitations of selecting significant domain data, applying the model to diverse domains and handling the complexity of the resultant BNs [9]. We derived the PRM based on the Control-Value theory of 'achievement emotions' [6], summarised in Table 1, along with considerations related to the goals of the *PlayPhysics* application. From this PRM were derived three Dynamic Bayesian Networks (DBNs), each corresponding to one of the types of emotion defined in [6]. It is important to note that whilst an appraisal of control or value is absent there is no emotion. The DBN corresponding to the '*outcome-prospective*' emotions is shown in Figure 1.

Each DBN was derived from analyzing the AEQ [19], e.g. one of the statements in the AEQ related to the emotion enjoyment is, "I am looking forward to learning a lot in this class", from this statement it was inferred that the student's level of performance, a cognitive factor, may be an indicator of positive, neutral or negative value and low, medium or high control. In addition, from analyzing the work in [18], we decided to focus only on motivational and cognitive variables in this stage of the research, since it was inferred that physiological variables can slightly increase the accuracy of the model. The work in [18] and [17] were used as a reference to identify observable variables and to define some dependencies with the motivational variables' effort, confidence and independence. The Conditional Probability Tables (CPTs) were set using common sense related to the learning context. To know the student's beliefs, attitudes and standards and to assess qualitatively and quantitatively

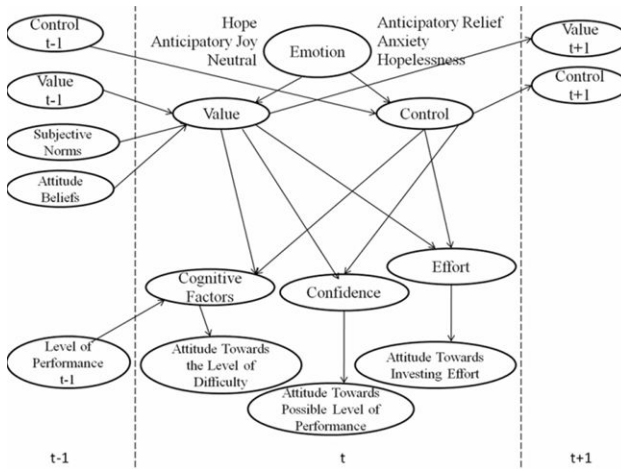


Fig. 1. Outcome-prospective emotions DBN

the learner's intentions, questions posed in game-dialogues are employed. These questions were derived using the 'Theory of Planned Behaviour' [34].

An evaluation of this affective model will be carried out with students of Physics at undergraduate level. The prototyping exercise, using a "Wizard-of-Oz" experiment [11] as a basis, will be performed with written material, which defines the dynamic of the game and the first game challenge. The student will be located in a Gesell dome. One lecturer will be in the same room providing assistance to solve the problem in the case the student needs it. Two lecturers will be behind the mirror, both making annotations of the observations related to the random variables. One of them will introduce the evidence acquired through these observations into 'Elvira' [35], a tool for implementing BNs, and will compare the results of the corresponding DBN with the emotion reported by the student. When the accuracy of the affective student model is significant, the model will be incorporated into *PlayPhysics*' architecture, *Olympia*. A preliminary evaluation of the prototyping material and the affective student employed for identifying the *outcome-prospective* emotions, discussed in section 5, was conducted with seven students at postgraduate level in the University of Ulster.

## 4 PlayPhysics Design and Implementation

An online survey was conducted with undergraduate students enrolled in an introductory course of Physics at Trinity College Dublin and Tecnológico de Monterrey, Mexico City, from March to December, 2009. According to identified requirements, the design of *PlayPhysics* was derived. Therefore, the topics that *PlayPhysics* is focused on teaching are circular movement, movement of rigid bodies, vectors and linear momentum, which are generally considered to be the most difficult topics on an introductory Physics course and involve Newton's laws of motion. *PlayPhysics*' storytelling scenario is a space adventure, where the student takes the role of an astronaut and is tasked with solving challenges using principles and knowledge of Physics. This section describes *PlayPhysics*' architecture, *Olympia*, the design and implementation

of *PlayPhysics*' first challenge and the definition of *PlayPhysics*' knowledge-base, i.e. cognitive model, related to this challenge.

### 4.1 PlayPhysics Architecture

The architecture of *PlayPhysics*, namely *Olympia* [12], has proven effective for teaching Physics at undergraduate level. *Olympia* is being modified in this research to incorporate the affective dimension and is shown in Figure 2. *Olympia* handles *dynamic* and *static interactive modules*.

Dynamic modules change over time while static modules remain the same. *Olympia* is a semi-open game learning environment [36], where learning goals guide the learner's interaction. *Olympia* includes an ITS. Events are selected by the interface analysis module, which sends them to the behaviour analysis module to be evaluated. The resultant evidence is communicated to the student model. The result of the inference is transferred to the tutor model and will be implemented using DDNs. The tutor model selects the action that will maximize learning or engagement. After receiving the action, the cognitive and affective modulators select the media according to the decisions taken by the planner. The presentation content manager modifies the world model and game mechanics module. Finally, the world model and the game mechanics influence the dynamic modules. To test the effectiveness of *PlayPhysics*, students at undergraduate level will be divided in control and experimental groups. Both groups will conduct a pre-test before learning the topics practiced with *PlayPhysics*.

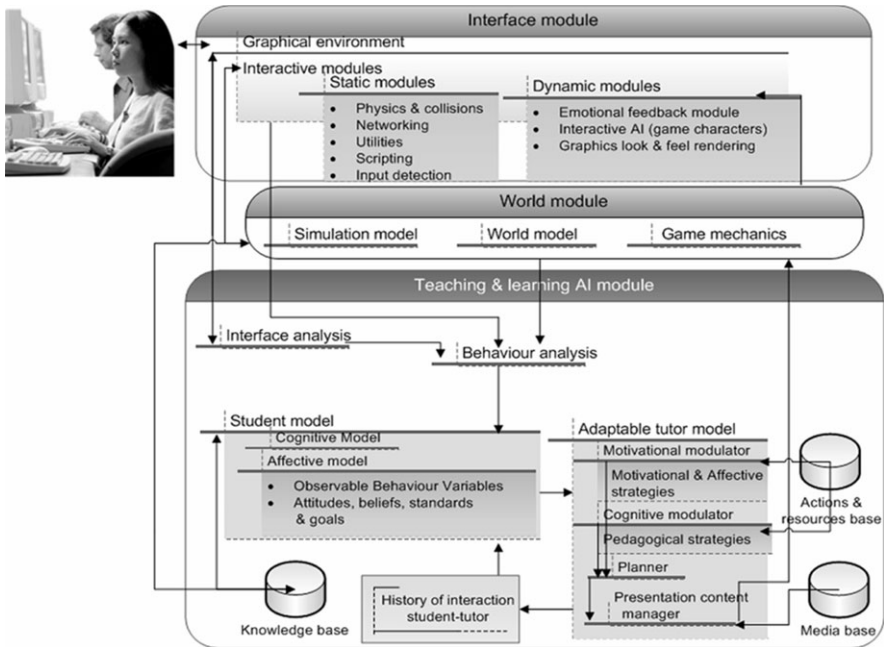


Fig. 2. Olympia Architecture



Then the experimental group students will interact with PlayPhysics and also attend their lectures, the control group will only attend the lectures. Both groups will conduct a post-test where the learning gains and learning efficiencies will be measured.

## 4.2 PlayPhysics' First Challenge Scenario Design and Implementation

The main goal of the game is to save Captain Richard Foster and re-establish control over the space station Athena. VNUS-2781, a super-computer, which controls Athena has been infected with a harmful virus and has attacked Athena's crew. Foster was the only person that could not escape. He was injured during the attack. To achieve the main goals, the student has to overcome challenges applying principles and concepts of Physics. The first challenge begins when NASA asks for the student's help. During this dialogue, *PlayPhysics* acquires information related to the student's attitudes, beliefs and standards through posed questions. This information and the result of a pre-test are used as evidence to reason about the learner's *outcome-prospective* emotions. In the first challenge, the astronaut has to dock his spaceship with Athena. Athena orbits around the sun and is located between Mars and Jupiter orbits and rotates with a constant angular velocity. The spaceship's initial velocity and location are randomly initialised to avoid triviality on the problem solving process. To preserve fuel, a minimal number of translational and rotational movements must be performed to dock with Athena. To achieve this goal the student must apply knowledge on linear and circular kinematics, Newton's laws for motion of particles and rigid bodies and vectors.

The spaceship is initially launched from Earth. First, the student has to slow and stop the spaceship. The spaceship stops somewhere in front of Athena's rotational axis using its front engines, i.e. the spaceship linearly decelerates. Second, the student must align the spaceship's longitudinal axis with Athena's longitudinal axis through turning on the upper and lower engines. Third, the spaceship's lateral engines are employed to achieve Athena's rotational velocity. Finally, the student enters slowly into Athena station though performing slow movements along its rotational axis. The student can explore the effects of modifying the spaceship's mass and rotational inertia relative to its three axes, e.g. longitudinal, zenithal and azimuthal. In addition, the forces and torques corresponding to spaceship's motors can be changed to modify the spaceship's angular acceleration and deceleration. The game-based learning environment is being implemented with the Unity Game Engine, Java Web, Elvira and Poser.

*PlayPhysics*' pedagogical actions will be defined through the suggestions of expert lecturers and implemented through Dynamic Decision Networks (DDNs). Game characters, colours and sounds will be mapped and displayed according to the emotions defined by Control-Value theory. *PlayPhysics*' learning effectiveness will be tested through dividing undergraduate students into control and experimental groups. Both groups will solve a pre-test before receiving the lecturers corresponding to the topics. The control and experimental groups will proceed to receive the lecturers. The experimental group will also interact with *PlayPhysics*. Finally, both groups will answer a post-test. The calculated learning gains and learning efficiencies between groups will be compared.

### 5 Preliminary Evaluation of the Affective Student Model

A pre-test, comprised of five questions related to the topics taught by *PlayPhysics*, and a questionnaire, comprising *PlayPhysics*' game dialogue, were applied to twenty eight students at undergraduate level from the Tecnológico de Monterrey (ITESM-CCM), Faculty of Engineering. The game dialogue questions correspond to the students' attitudes, beliefs and standards with regard to Physics. The previous experience of solving the pre-test was also considered whilst answering these questions. The evidence obtained through both questionnaires was propagated into the *outcome-prospective* emotions DBN. The main aim of this evaluation was to identify possible deficiencies in the prototyping material and to evaluate the *outcome-prospective* DBN accuracy at reasoning about the students' emotions. Results, obtained through comparing the inferred emotion by *PlayPhysics*' affective student model with the students' reported emotion, showed an accuracy of 60.71% (Figure 3).

The results were promising, but the model still needs to be evaluated through a larger population of participants. As an example, the authors in [19] validated the AEQ with the participation of 389 students. Students signalled that to enhance the understanding of the questions comprising the pre-test, an explanation of some terms comprising formulae and diagrams supporting the questions can be included. In addition, it was noted that some students did not know how to classify the emotion that they were feeling. Therefore, in a future version, examples corresponding to the emotions will be included. The probabilities in the Conditional Probability Tables (CPTs) set using common-sense related to the learning context may also be influencing the results. In addition, each student will be evaluated separately in the Gesell dome, since it was noted that when the students are sharing the same task in the same location, some of them have a tendency of behaving in a competitive way. For example, two students reported anxiety, even when they achieved a performance of over 70% in their pre-tests. They thought that their performances will be published and made available to everyone who took the pre-test.

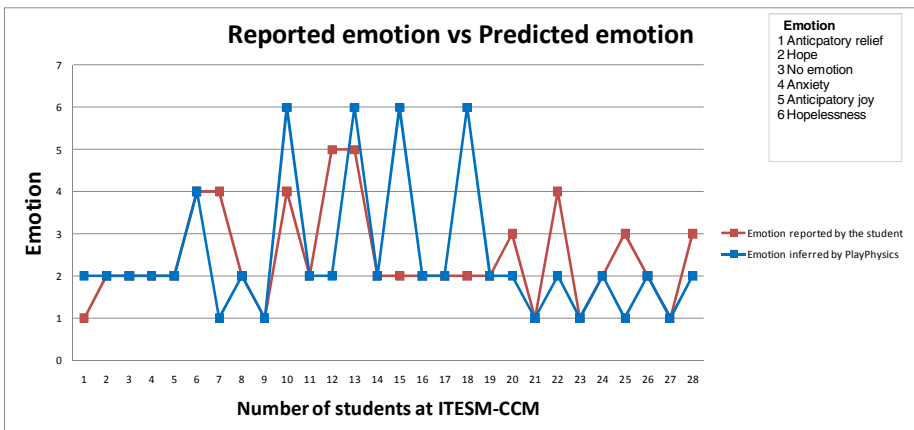


Fig. 3. Students' reported emotion and PlayPhysics' inferred emotion

## 6 Conclusion and Future Work

The state of the art related to research challenges in identifying and influencing the learner's emotion was reviewed. As of today, there is no system that reasons effectively about the learner's emotion using a combination of cognitive and motivational variables. Therefore we focus on creating such affective student model. The model was designed using the Control-Value theory as a basis and was implemented using a PRMs approach and DBNs. The advantages of modulating the output to influence the learner's affective state were noted. Game elements, e.g. sounds, visuals and colours, serve meaningful purposes and will be employed to reduce the learning curve and set a mood. A preliminary evaluation showed that the affective student model has an effectiveness of 60.71% when reasoning about the students' *outcome-prospective* emotions. However, the complete affective student model will be tested on a larger population of students at undergraduate level, enrolled on an introductory Physics course. Deficiencies, which were detected in the prototyping material, will be modified. In addition, the effectiveness of *PlayPhysics* for teaching will be evaluated through the comparison of learning gains and efficiencies.

## References

1. Oblinger, D.G.: The Next Generation of Educational Engagement. *Interactive Media in Education* (8), 1–18 (2004)
2. The ICE House Project Group: Teaching for Innovation, Creativity and Enterprise (2009), [http://www.ice-house.info/Resources/ICE\\_House\\_Project\\_1st\\_Briefing\\_Paper.pdf](http://www.ice-house.info/Resources/ICE_House_Project_1st_Briefing_Paper.pdf)
3. Squire, K.: Video Games in Education. *International Journal of Intelligent Simulations and Gaming* 2(1), 49–62 (2003)
4. Conati, C., Maclaren, H.: Empirically Building and Evaluating a Probabilistic Model of User Affect. *User Modeling and User-Adapted Interaction* 19(3), 267–303 (2009)
5. Rebolledo-Mendez, G., Du Boulay, B., Luckin, R.: Motivating the Learner: An Empirical Evaluation. In: Ikeda, M., Ashley, K., Chan, T.-W. (eds.) *ITS 2006*. LNCS, vol. 4053, pp. 545–554. Springer, Heidelberg (2006)
6. Pekrun, R., Frenzel, A.C., Goetz, T., Perry, R.P.: The Control Value Theory of Achievement Emotions. An integrative Approach to Emotions in Education. In: Shutz, P.A., Pekrun, R. (eds.) *Emotion in Education*, pp. 13–36. Elsevier, London (2007)
7. Picard, R.W., Papert, S., Bender, W., Blumberg, B., Breazeal, C., Cavallo, D., Machover, T., Resnick, M., Roy, D., Strohecker, C.: Affective learning –A Manifesto. *BT Technology Journal* 22(4), 253–269 (2004)
8. Sykes, J.: Affective Gaming: Advancing the Argument for Game-Based Learning. In: Pivec, M. (ed.) *Affective and Emotional Aspects of Human-Computer Interaction*, pp. 3–7. IOS Press, Netherlands (2006)
9. Sucar, L.E., Noguez, J.: Student Modeling. In: Pourret, O., Naïm, P., Marcot, B. (eds.) *Bayesian Networks: A Practical Guide to Applications*, pp. 173–185. J. Wiley & Sons, West Sussex (2008)
10. Du Boulay, B., Luckin, R.: Modelling Human Teaching Tactics and Strategies for Tutoring Systems. *International Journal of Artificial Intelligence in Education* 12, 235–256 (2001)

11. Höök, K.: User-Centred Design and Evaluation of Affective Interfaces. In: Ruttkoy, Z., Pelachaud, C. (eds.) *From Brows to Trust: Evaluating Embodied Conversational Agents*, pp. 127–160. Springer, Netherlands (2005)
12. Muñoz, K., Noguez, J., Mc Kevitt, P., Neri, L., Robledo-Rella, V., Lunney, T.: Adding features of educational games for teaching Physics. In: 39th IEEE International Conference Frontiers in Education, pp. M2E-1–M2E-6. IEEE Press, USA (2009)
13. Lepper, M.R., Woolverton, M., Mumme, D.L.: Motivational Techniques of Expert Human Tutors: Lessons for the Design of Computer Based Tutors. In: Lajoie, S.P., Derry, S.J. (eds.) *Computers as Cognitive Tools*, pp. 75–105. Lawrence Erlbaum Associates, Mahwah (1993)
14. D’Mello, S.K., Craig, S.D., Witherspoon, A., McDaniel, B.T., Graesser, A.C.: Automatic Detection of Learner’s Affect from Conversational Cues. *User modelling and User-Adapted interaction* 8(1-2), 45–80 (2008)
15. Jaques, P.A., Vicari, R.M.: A BDI Approach to Infer Student’s Emotions in an Intelligent Learning Environment. *Journal of Computers and Education* 49(2), 360–384 (2007)
16. Ortony, A., Clore, G.L., Collins, A.: *The Cognitive Structure of Emotions*. Cambridge University Press, NY (1990)
17. Del Soldato, T., Du Boulay, B.: Implementation of motivational tactics in tutoring systems. *Journal of Artificial Intelligence in Education* 6(4), 337–378 (1995)
18. McQuiggan, S.W., Mott, B.W., Lester, J.C.: Modeling Self-efficacy in Intelligent Tutoring Systems: An inductive approach. *User Modelling and User-Adapted Interaction* 18, 81–93 (2008)
19. Pekrun, R., Goetz, T., Perry, R.P.: *Achievement Emotions Questionnaire (AEQ). User’s manual*. Unpublished manuscript, University of Munich, Munich (2005)
20. Johnson, W.L., Rickel, J.W., Lester, J.C.: Animated pedagogical agents: face to face interaction in interactive learning environments. *International Journal of Artificial Intelligence in Education* 11(1), 47–78 (2000)
21. Stone, B.A., Lester, J.C.: Dynamically sequencing an animated pedagogical agent. In: *Proceedings of the 13th National Conference on Artificial Intelligence*, pp. 424–431. The MIT Press, Portland (1996)
22. Bates, J., Loyall, A., Reilly, W.: Integrating reactivity, goals and emotion in a broad agent. *CiteSeerX* (1992), <http://www.cs.cmu.edu/afs/cs/project/oz/web/papers/CMU-CS-92-142.ps.gz>
23. Lester, J.C., Voerman, J.L., Towns, S.G., Callaway, C.B.: Diectic believability: Coordinating gesture, locomotion and speech in life-like pedagogical agents. *Applied Artificial Intelligence* 13, 383–414 (1999)
24. Cassell, J., Högni Vilhjálmsson, H., Bickmore, T.: BEAT: The Behaviour Expression Animation Toolkit. In: Pocock, L. (ed.) *Proceedings of the 28th Annual Conference on Computer Graphics and Interactive Techniques*, pp. 477–486. ACM Press, Los Angeles (2001)
25. Streit, M., Batliner, A., Portele, T.: Cognitive-Based Interpretation of Emotions in a Multimodal Dialog System. In: Carbonell, J.G., Siekmann, J. (eds.) *ADS 2004. LNCS (LNAI)*, vol. 3068, pp. 65–76. Springer, Heidelberg (2004)
26. Mateas, M.: *An Oz-Centric Review of Interactive Drama and Believable Agents*. Unpublished manuscript, School of Computer Science, Carnegie Mellon University (1997)
27. Dias, J., Paiva, A., Vala, M., Aylett, R., Woods, S., Zoll, C., Hall, L.: Empathic characters in computer-based personal and social education. In: Pivec, M. (ed.) *Affective and Emotional Aspects of Human-Computer Interaction*, pp. 246–254. IOS Press, Netherlands (2006)

28. Collins, K.: *Game sound: an introduction to the history, theory and practice of video game music and sound design*. MIT Press, Cambridge (2008)
29. Malone, T.W.: *Toward a Theory of Intrinsically Motivating Instruction*. *Cognitive Science* 5(4), 333–369 (1981)
30. Kaya, N., Epps, H.H., Hall, D.: *Relationship between color and emotion: a study of college students*. *College Student Journal*, 396–405 (2004)
31. Alsmeyer, M., Luckin, R., Good, J.: *Developing a novel interface for capturing self-reports of affect*. In: *Proceedings of the CHI 2008: Conference on Human Factors in Computing Systems*, pp. 2883–2888. ACM Press, Florence (2008)
32. Razek, M.A., Chaffar, S., Frasson, C., Ochs, M.: *Using machine learning techniques to recognize emotions for online learning environments*. In: Pivec, M. (ed.) *Affective and Emotional Aspects of Human-Computer Interaction*, pp. 255–265. IOS Press, Netherlands (2006)
33. Nijdam, N.A.: *Mapping emotion to color* (2005),  
<http://hmi.ewi.utwente.nl/verslagen/capita-selecta/CS-Nijdam-Niels.pdf>
34. Francis, J.J., Eccles, M.P., Johnston, M., Walker, A., Grimshaw, J., Foy, R., Kaner, E.F.S., Smith, L., Bonetti, D.: *Constructing Questionnaires based on the Theory of Planned Behaviour: a manual for Health Services Researchers*. Unpublished manuscript, University of Newcastle, Newcastle, UK (2004)
35. Díez, J.: *The Elvira Project* (2005),  
<http://www.ia.uned.es/~elvira/index-en.html>
36. Bunt, A., Conati, C.: *Probabilistic Student Modeling to Improve Exploratory Behavior*. *Journal of User Modeling and User-Adapted Interaction* 13(3), 269–309 (2003)

# A SOM-Based Technique for a User-Centric Content Extraction and Classification of Web 2.0 with a Special Consideration of Security Aspects

Amirreza Tahamtan<sup>1</sup>, Amin Anjomshoaa<sup>1</sup>, Edgar Weippl<sup>1,2</sup>, and A. Min Tjoa<sup>1</sup>

<sup>1</sup> Vienna University of Technology, Dept. of Software Technology & Interactive Systems, Information & Software Engineering Group  
{tahamtan, anjomshoaa, amin}@ifs.tuwien.ac.at

<sup>2</sup> Secure Business Austria, Favoritenstrae 16 - 2nd floor, 1040 Vienna, Austria  
eweippl@securityresearch.at

**Abstract.** Web 2.0 is much more than adding a nice facade to old web applications rather it is a new way of thinking about software architecture of Rich Internet Applications (RIA). In comparison to traditional web applications, the application logic of modern Web 2.0 applications tends to push the interactive user interface tasks to the client side. The client components on the other hand negotiate with remote services that deal with user events. The user should be assisted in different scenarios in order to use the existing platforms, share the resources with other users and improve his security. In this paper we present a user-centered content extraction and classification method based on self-organizing maps (SOM) as well as a prototype for provided content on Web 2.0. The extracted and classified data serves as a basis for above mentioned scenarios.

**Keywords:** Web 2.0, Self-organizing maps, Extraction, Classification, Security.

## 1 Introduction

Web 2.0 makes a better use of the client and at the same time pushes the SOA paradigm [23] to its limits. Web 2.0 envisions building collective intelligence and mashed up functionality [4] based on web services. Web 2.0 Users should be supported in three different scenarios:

**Assistive services:** The specific Web 2.0 contents should provide assistance for users who create similar contents. By analyzing existing contents, some templates and structures should be established and suggested to other users for common contents. It is important to note that the assistive services are not allowed to share sensitive data with other users.

**Resource sharing:** The data sharing on Web 2.0 is decided by the content owner and there is no holistic solution to avoid an unwanted information disclosure. There is an ever growing need for intelligent sharing of information based on the context.

**Self-monitoring of trust level:** The data contributed by users on Web 2.0 is a source of judgement about individual/organizational behaviors and attitude. This includes the membership in social networks, user groups, contributions on Wikipedia, blog entries, shared videos and pictures and virtual games. In some cases these inferences are not correct and the individuals and organizations have no means to prevent false judgements.

Whatever the intended scenario, the Web 2.0 solutions should satisfy the following requirements:

- The platform should be generic and scalable. Scalability is important since the information on the web is being rapidly increased and the platform should cope with huge amount of data.
- Definition of standard methods for analysis of Web 2.0 documents according to specific resource data models. The result of content analysis should provide the feed to data sharing policies.
- Supporting the user for the development of Web 2.0 content by formal identification of relevant information by means of content analysis results, user behavior and domain ontology.
- Relating the web items that are published in different languages and mapping them to the same ontology in order to make better inferences.

The prerequisite for reaching the above mentioned goals of assisting the user, is automatic extraction and classification of data. For example if a user wants to use available templates on the Web for brainstorming, already created templates should be extracted, ranked and subsequently suggested to the user, as e.g. proposed in [5]. In other words, available templates must be extracted and classified. In another scenario if a user wants to check and monitor his contributions on the Web, e.g. for movie rating sites his entries should be extracted and classified according to the topic. Because of the wide range of platforms on the Web 2.0 from social networks to movie rating sites, performing such tasks manually is very tedious and almost impossible.

This paper provides an overview on the overall approach of our project Secure 2.0 (Securing the Information Sharing on Web) 2.0 [3]. The main contribution of this paper is the introduction of an approach and prototype based on self-organizing maps (SOM) that extracts and classifies the provided content on Web 2.0. The results can be used for assisting the users in the above mentioned scenarios and serves as a basis for these aims. The presented approach is user-centered in the sense that the user himself can assess his web presence before being evaluated by other persons or authorities.

## 2 Security on Web 2.0

The need for usable and trusted privacy and security is a critical area in the management of Web 2.0 information. This goal demands not only efficient security and privacy policies but also requires improvements of the usability of security aspects.

The disclosure of personal/organizational information on Web 2.0 has created new security and privacy challenges. Designing transparent, usable systems in support of personal privacy, security and trust includes everything from understanding the intended use of a Web 2.0 system to users' tasks and goals as well as the contexts in which the users use the system for information sharing purposes.

Web 2.0 security concerns are divided into two basic categories: physical security and semantic security. The former aims to cover issues such as secure and trustworthy data exchange. This group of security concerns can benefit from existing methodologies of Web 1.0. The semantic security handles the information sharing on a higher level by exploiting the Semantic Web technologies in order to describe the shared knowledge in a computer-processable way. As a result the shared information can be combined with personal/organizational policies to protect the information in a collaborative environment like Web 2.0.

Obviously new security and privacy schemas are required to cover the requirements of Web 2.0 applications which are being raised due to the following reasons:

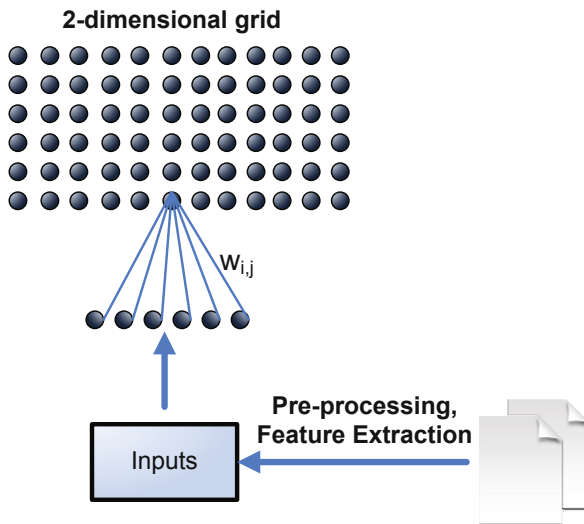
- Web services are the building blocks of Web 2.0 applications and by liberating web services from organizational environments, it is necessary to have appropriate information disclosure and information usage policies.
- Web 2.0 has made the content creation much easier and as a result a huge amount of data is constantly created. The volume of data on the web is doubled since the emergence of Web 2.0 technologies. Data mining of user generated entities and extracting knowledge and information patterns is a new threat to privacy of individuals.
- The Web 2.0 architecture is tending to utilize the client side processing power. Hence, Web 2.0 can be used intelligently for better integration of user data with global business processes. In other words the "user desktop" can interact with the real world processes and provide the requested data without human interaction. Before such dreams can come true, we need an efficient mechanism to define users' security and privacy policies.

### 3 Self-Organizing Maps

Self-organizing maps (SOM) or self-organizing feature maps, also sometimes called Kohonen maps, have been introduced by Teuvo Kohonen in [12]. SOM belongs to the family of artificial neural networks and uses unsupervised learning. It can be used for reducing the dimensionality of the input by mapping the input onto a low-dimensional (usually two dimensional) grid. Each input object is represented as an N-dimensional vector. Each dimension of the vector is called a feature. Each node in the grid is assigned a weight vector which is again represented by an N-dimensional vector. Components of a weight vector are assigned at initialization time random values. The SOM's learning process is as follows: a node in the grid with the minimum distance to a given input is chosen. This node is called the winning node. The components of the weight vector of the



winning node are adjusted such that the distance with the input vector becomes smaller. The components of the weight vectors of the neighboring nodes of the winning node are as well adjusted. This cycle is iterated until it converges, i.e. no more adjustments are performed. Figure 1 depicts an architecture based on SOM for information retrieval from documents. For more details on analyzing textual data we refer to subsection 3.1 and section 4. The main characteristic of SOM is preserving topology, i.e. the neighborhood and distance relationship between input data is preserved and by mapping becomes explicit. SOM maps more frequent input data onto larger domains compared to less frequent input data. Several Authors have proposed different variants of SOM and several algorithm for it has been developed, e.g. [9,8].



**Fig. 1.** A SOM-Based approach for information retrieval from documents

SOM has many applications in different domains. An overview on SOM-related literature can be found in [10,18].

### 3.1 SOM-Based Text Analysis

In this work we use SOM for automatic clustering of high-dimensional data. The SOMs can be visualized and the distance between concepts depicts their similarity with regards to some predefined features. A typical SOM algorithm for classification of text based items can be summarized as follows [6]:

1. Initialize input nodes, output nodes, and weights: Use the top (most frequently occurring)  $N$  terms as the input vector and create a two-dimensional map (grid) of  $M$  output nodes. Initialize weights  $w_{ij}$  from  $N$  input nodes to  $M$  output nodes to small random values.

2. Present each document in order: Describe each document as an input vector of  $N$  coordinates. Set a coordinate to 1 if the document has the corresponding term and to 0 if there is no such a term.
3. Compute distance to all nodes: Compute Euclidean distance  $d_j$  between the input vector and each output node  $j$ :

$$d_j = \sum_{i=0}^{N-1} (x_i(t) - w_{ij}(t))^2$$

where  $x_i(t)$  can be 1 or 0 depending on the presence of  $i$ -th term in the document presented at time  $t$ . Here,  $w_{ij}$  is the vector representing position of the map node  $j$  in the document vector space. From a neural net perspective, it can also be interpreted as the weight from input node  $i$  to the output node  $j$ .

4. Select winning node  $j^*$  and update weights to node  $j^*$  and its neighbors: Select winning node  $j^*$ , which produces minimum  $d_j$ . Update weights to nodes  $j^*$  and its neighbors to reduce the distances between them and the input vector  $x_i(t)$ :

$$w_{ij}(t+1) = w_{ij}(t) + \eta(t)(x_i(t+1) - w_{ij}(t))$$

After such updates, nodes in the neighborhood of  $j^*$  become more similar to the input vector  $x_i(t)$ . Here,  $\eta(t)$  is an error-adjusting coefficient ( $0 < \eta(t) < 1$ ) that decreases over time.

5. After the network is trained through repeated presentations of all documents, assign a term to each output node by choosing the one corresponding to the largest weight (winning term). Neighboring nodes which contain the same winning terms are merged to form a concept/topic region (group). Similarly, submit each document as input to the trained network again and assign it to a particular concept in the map.

Figure 2 shows an example of a self-organizing map with clusters and sub-clusters.

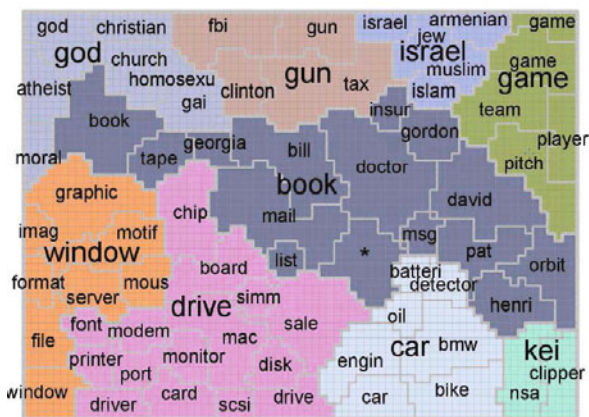
## 4 The Proposed Approach

There are a handful of techniques and algorithms for text analysis and information retrieval (IR) purposes. It goes without saying that the rudimentary methods such as removal of stop words and stemming are not enough.

To analyze textual data with a self-organizing map, the following steps must be performed:

**Data Extraction:** The first step toward using Web 2.0 contents is to analyze and extract that data. In this context the Web 2.0 API of the systems under study will provide the required feeds for the text analysis component.

**Pre-processing:** This step includes applying stemming algorithms, removing stop words, format conversion, etc.



**Fig. 2.** Self-organizing map with cluster and sub-clusters, image from [22]

**Feature Extraction:** Feature extraction refers to description of data characteristics. Note that self-organizing maps are only capable of handling numerical data. Hence features need to be presented in a numerical form. Textual data can be represented by different approaches such as bag-of-words [28], phrase detection [11] and Latent Semantic Indexing [7]. We use a binary presentation of features, i.e. components of the input vectors, which represent the words contained in the document are assigned 1 if this word is contained in the document and 0 otherwise. The features of a document will be presence or absence of words in this document. For example consider the following entries on Twitter. Twit 1: "The Guardian newspaper has announced it will support the Liberal Democrats" and Twit 2: "A woman brought you into this world, so you have no right to disrespect one". Their extracted vector is depicted in the figure 3. The vector shows the results after the pre-processing step, e.g. removal of stop words, stemming, etc. That is why only a subset of the words are included in the vector. Note that this a simplified version. In our experiments we consider the correct sense of the word in the context using the WordNet [16] and augment the vector with this information as described in subsection 3.1.

**Training:** After the input data has been prepared, it can be used to train the self-organizing maps. There is no general rule about the map size. However, the number of data items must be sufficiently large enough compared to the map size. According to [13], eleven training samples per node is just about sufficient.

**Visualization:** The maps can be visualized using a variety of methods such as: Vector Activity Histogram, Class Visualization [15], Component Planes [27], Vector Fields [20], Hit Histogram, Metro Map [17], Minimum Spanning Tree, Neighborhood Graph [21], Smoothed Data Histograms [19], Sky Metaphor Visualization [14], U-Matrix [26], D-Matrix, P-Matrix [24] and  $U^*$ -Matrix [25].

	guardian	newspaper	announce	support	liberal	democrat	woman	bring	world	right	disrespect
Twit 1:	1	1	1	1	1	1	0	0	0	0	0
Twit 2:	0	0	0	0	0	0	1	1	1	1	1

**Fig. 3.** An example of an input vector

Furthermore as an additional task we have to take the extra information of Web 2.0 entries into consideration throughout the document processing phase. As mentioned before, the Web 2.0 entries might have an additional semi-structured information such as tags, attachments, relations, etc. This information can play an important role in the item analysis of Web 2.0 entries and should be extracted and included in the input vector of the SOM algorithm. At the end of the SOM process we end up with a group of data points that are merged together and form clusters. The clusters are labeled and this labeling might be further improved in the next step for creating the ontologies. An important matter in this step is consideration of security and privacy issues. After the analysis of documents the sensitive content of the nodes should be removed. The use of the extracted ontology together with context and security ontology make this possible. Security issues are discussed in more detail in subsection 4.1.

#### 4.1 Security and Privacy Solutions

The text analysis process of the previous step is focused on a domain-independent, statistical analysis of the text. This should be combined with context information. The statistically derived "implicit semantics" of Web 2.0 entries should be annotated and aligned with "explicit semantics" which is based on formal ontology and domain knowledge. Ontology and semantic metadata also play a critical role in combining the existing knowledge with application context by scoring and ranking the fitting candidate information. Moreover the "explicit semantics" is used to bridge the gap between content, users and policies via their relevant ontologies. The "implicit semantics" of our SOM-approach may play the role of an indicator to specify the priority of required domain ontologies. As a result the Web 2.0 items plus domain ontology provide an elaborated set of information for improving inference and query answering processes. Ontologies are used as the basis for providing assistive-services and information sharing. Also in self-monitoring use case the ontology plays a crucial role to bind the information from different resources together.

Part of the alignment process can be done automatically by selecting the appropriate ontology from available domain ontologies and annotating the Web 2.0 items with ontology concepts. The major applications of ontology alignment are as follows:

- Formal domain ontology together with security and privacy ontologies make it possible to identify and anonymize the sensitive data. The result can be safely shared with other users or be reused via an assistive service as a template.

- Domain ontologies can be also combined with user policy to restrict the information sharing and apply necessary filters.
- Ontology alignment process can be used as an input for ontology engineers to enrich the domain ontologies based on common data structures.

The Web 2.0 items via their relevant domain ontologies is connected to other information resources for more elaborated tasks such as reasoning or complex queries. As soon as we arrive in the ontology level, the solutions can benefit from the concrete research works about semantic security and privacy. Figure 4 depicts the overall solution and the important role that ontologies play in connecting the distributed knowledge domains

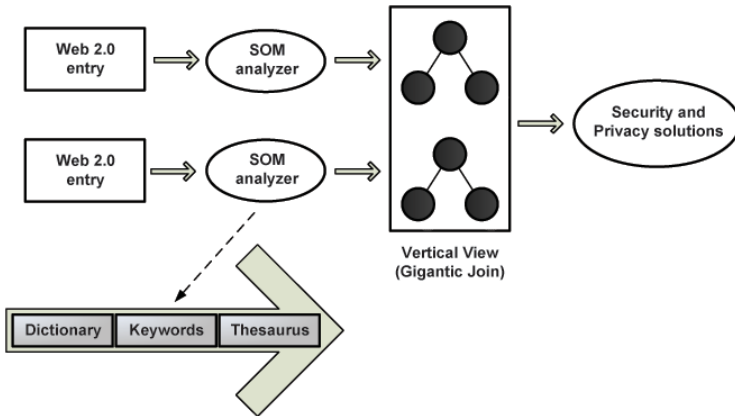
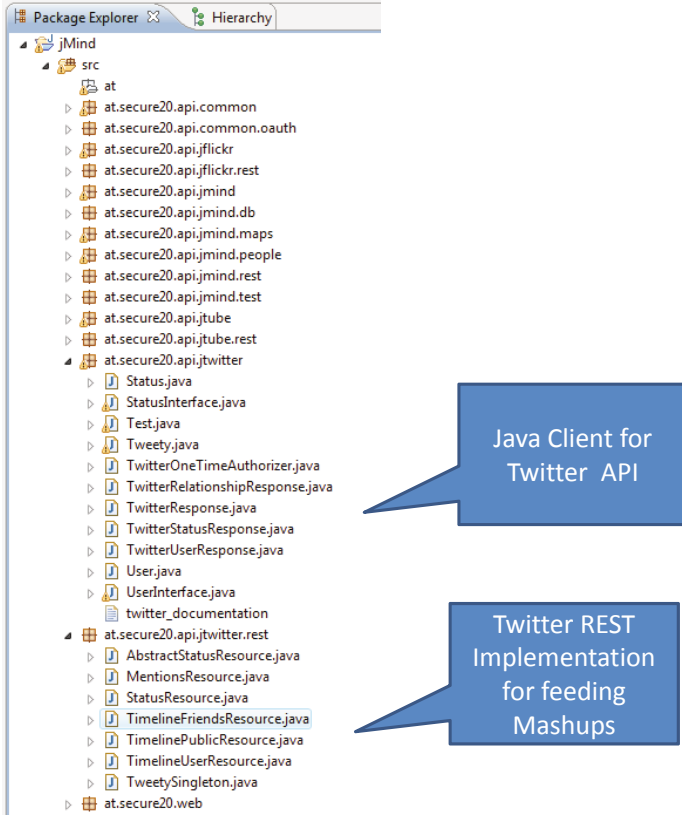


Fig. 4. overall solution and ontological join of knowledge domains

## 5 Experiments and Prototypical Implementation

Before starting with knowledge extraction and advanced text processing techniques we setup a data resource for our experiments. For this purpose three major Web 2.0 platforms: YouTube, Flickr, and Twitter, plus MindMeister [1] were selected as the basic data resources. Java components that use the corresponding REST APIs of these platforms for extracting data items and storing them locally is implemented. Figure 5 shows a screen shot of this component. The extracted data can be again used as services. It is important to note that the extractor also plays an important role in the runtime system and will be used to create temporary data sources for user-generated scenarios on the fly. Furthermore a prototype for disambiguation and annotation of mind map content based on WordNet dictionary is implemented. The disambiguation results are then used to annotate the text with the correct sense of the word which is necessary before starting up with SOM for getting better results, i.e. to improve the quality of SOMs and decrease the error function in terms of misclassification. After feature extraction from documents, the SOM analysis is performed using the Java SOMToolbox [2].



**Fig. 5.** Data extractor and feed components

Figure 6 illustrates an experiment on Twitter entries. This experiment includes 1754 twits (number of vectors) and 1282 words after removal of stop words (number of features) on a  $11 \times 11$  map. Due to visibility reasons only a part of the map is visualized here. The numbers on the map show how many twits are about the same topic and therefore are mapped onto the same region. The black circles identify the twits that maybe problematic or about inappropriate topics. Using this approach the user can identify that a person he is following on the twitter is posting entries about topics which he does not want to be linked with, e.g. hate literature, and the user may want to remove him from his list. Identification of out of favor topics or entries are user’s responsibility. For example employee of a company may not want to be linked to positive posts on competitors’ products whilst other users have no problem with that. As users have different legal, ethical, personal, religious and job ethics, we have left this decision up to the user.

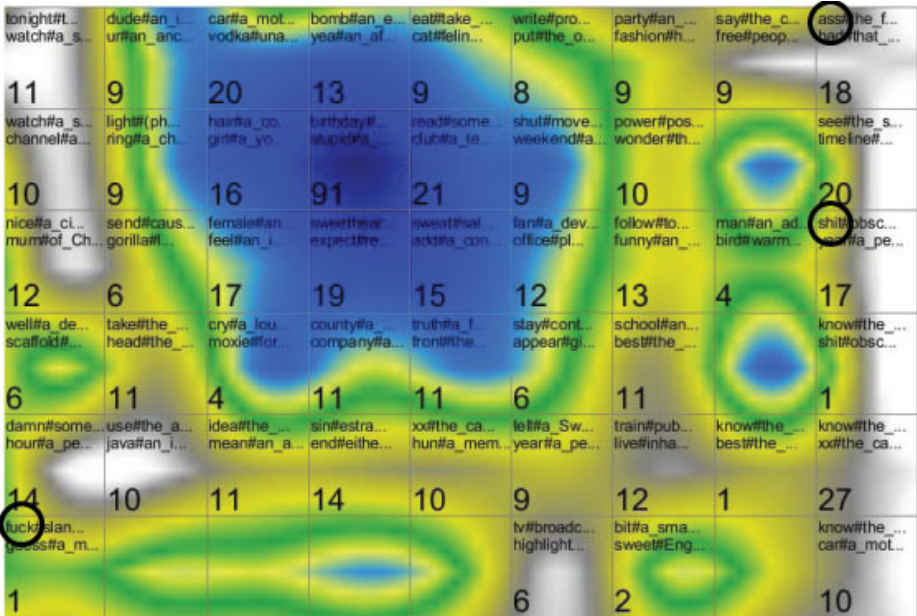


Fig. 6. Smoothed data histogram visualization of Twitter entries

## 6 Conclusions and Future Work

Data extraction and classification can be used to in many ways to assist the user in different scenarios. In this paper we have presented a generic approach for extraction and classification of data in a user-centered way. In other words, the user himself can monitor his web presence before being evaluated by others. We use SOMs for classification and categorization of textual data. The provided results are parts of the ongoing project secure 2.0.

We plan to use Mashups to create situational solutions of our platform. This decision has two main benefits: first of all mashups will enable us to setup and test different approaches and service composition variants. Second, the end user of system who is not an IT expert can easily put the services together and create a customized solution that meets his requirements. There are also some other factors that will be studied and tested. For example the weight selection and the number of clusters are two major parameters in using SOM for clustering purposes. The first experiments show that we need a method to decrease and merge smaller clusters while preserving the topology of data. Inappropriate selection of clusters will result in huge number of clusters that will be computationally intensive. In this regard we are planning to benefit from other clustering methods such as K-means clustering. Another question is the optimal number of iterations in the training phase, the learning rate and the appropriate size of the 2-dimensional grid. Another problem that we are dealing with is the huge

amount of data that SOM should handle. In this regard we examine how to enhance the capabilities of the suggested algorithms for self-organizing maps such as scalable self-organizing maps to cater for the significant amount of data and how we can use the characteristics of the data present on Web 2.0 to reduce the computational complexity.

**Acknowledgments.** This work is supported by the Austrian FIT-IT project Secure 2.0.

## References

1. <http://www.mindmeister.com/>
2. <http://www.ifs.tuwien.ac.at/dm/somtoolbox/index.html>
3. Secure 2.0 - securing the information sharing on web 2.0, <http://www.ifs.tuwien.ac.at/node/6570>
4. Anjomshoaa, A.: Integration of Personal Services into Global Business. PhD thesis, Vienna University of Technology (2009)
5. Anjomshoaa, A., Sao, K.V., Tjoa, A.M., Weippl, E., Hollauf, M.: Context oriented analysis of web 2.0 social network contents - mindmeister use-case. In: Proc. of the 2nd Asian Conference on Intelligent Information and Database Systems (2010)
6. Chen, H., Schuffels, C., Orwig, R.: Internet categorization and search: a machine learning approach. *Journal of Visual Communications and Image Representation* 7(1), 88–102 (1996)
7. Deerwester, S., Dumais, S.T., Furnas, G.W., Landauer, T.K., Harshman, R.: Indexing by latent semantic analysis. *Journal of the American Society For Information Science* 41, 391–407 (1990)
8. Fritzke, B.: Growing cell structures—a self-organizing network for unsupervised and supervised learning. *Neural Networks* 7(9), 1441–1460 (1994)
9. Kangas, J.A., Kohonen, T., Laaksonen, J.T.: Variants of self organizing feature maps. *IEEE Transactions on Neural Networks* 1(1), 93–99 (1990)
10. Kaski, S., Kangas, J., Kohonen, T.: Bibliography of self-organizing map (som) papers: 1981-1997. *Neural Computing Surveys* 1(3-4), 1–176 (1998)
11. Kawahara, T., Lee, C.H., Juang, B.H.: Combining key-phrase detection and subword-based verification for flexible speech understanding. In: Proc. of the International Conference on Acoustic, Speech, Signal Processing (1997)
12. Kohonen, T.: The self-organizing map. *Proc. IEEE* 78, 1464–1480 (1990)
13. Kohonen, T., Somervuo, P.: Self-organizing maps of symbol strings. *Neurocomputing* 21(1-3), 19–30 (1998)
14. Latif, K., Mayer, R.: Sky-metaphor visualisation for self-organising maps. In: Proc. of the 7th International Conference on Knowledge Management (2007)
15. Merkl, D., Rauber, A.: Alternative ways for cluster visualization in self-organizing maps. In: Proc. of the Workshop on Self-Organizing Maps (1997)
16. Miller, G.A.: Wordnet: a lexical database for english. *Communications of the ACM* 38(11), 39–41 (1995)
17. Neumayer, R., Mayer, R., Rauber, A.: Component selection for the metro visualisation of the som. In: Proc. of the 6th International Workshop on Self-Organizing Maps (2007)
18. Oja, M., Kaski, S., Kohonen, T.: Bibliography of self-organizing map (som) papers: 1998-2001 addendum. *Neural Computing Surveys* 3, 1–156 (2002)



19. Pampalk, E., Rauber, A., Merkl, D.: Using smoothed data histograms for cluster visualization in self-organizing maps. In: Dorrnsoro, J.R. (ed.) ICANN 2002. LNCS, vol. 2415, p. 871. Springer, Heidelberg (2002)
20. Poelzlbauer, G., Dittenbach, M., Rauber, A.: Advanced visualization of self-organizing maps with vector fields. *Neural Networks* 19(6-7), 911–922 (2006)
21. Poelzlbauer, G., Rauber, A., Dittenbach, M.: Advanced visualization techniques for self-organizing maps with graph-based methods. In: Proc. of the Second International Symposium on Neural Networks (2005)
22. Roiger, A.: Analyzing, labeling and interacting with soms for knowledge management. Master's thesis, Vienna University of Technology (2007)
23. Tahamtan, A.: Modeling and Verification of Web Service Composition Based Interorganizational Workflows. PhD thesis, University of Vienna (2009)
24. Ultsch, A.: Maps for the visualization of high-dimensional data spaces. In: Proc. of the Workshop on Self-Organizing Maps (2003)
25. Ultsch, A.: U\*-matrix: a tool to visualize clusters in high dimensional data. Technical Report Technical Report No. 36, Dept. of Mathematics and Computer Science, University of Marburg, Germany (2003)
26. Ultsch, A., Siemon, H.P.: Kohonen's self-organizing feature maps for exploratory data analysis. In: Proc. of the International Neural Network Conference (1990)
27. Vesanto, J., Ahola, J.: Hunting for correlations in data using the self-organizing map. In: Proc. of the International ICSC Congress on Computational Intelligence Methods and Applications (1999)
28. Wallach, H.M.: Topic modeling; beyond bag of words. In: Procs. of the International Conference on Machine Learning (2006)

# Modularizing Spatial Ontologies for Assisted Living Systems

Joana Hois

Spatial Cognition Research Center SFB/TR8  
University of Bremen  
joana@informatik.uni-bremen.de

**Abstract.** Assisted living systems are intended to support daily-life activities in user homes by automatizing and monitoring behavior of the environment while interacting with the user in a non-intrusive way. The knowledge base of such systems therefore has to define thematically different aspects of the environment mostly related to space, such as basic spatial floor plan information, pieces of technical equipment in the environment and their functions and spatial ranges, activities users can perform, entities that occur in the environment, etc. In this paper, we present thematically different ontologies, each of which describing environmental aspects from a particular perspective. The resulting modular structure allows the selection of application-specific ontologies as necessary. This hides information and reduces complexity in terms of the represented spatial knowledge and reasoning practicability. We motivate and present the different spatial ontologies applied to an ambient assisted living application.

**Keywords:** Ontologies, Knowledge engineering, Conceptual modeling in knowledge-based systems, Knowledge-based systems in life sciences.

## 1 Introduction

Ontologies provide tools for organizing and contextualizing knowledge [4]. They are widely used in different fields as a method for making explicit what is already known implicitly. Their terminology is supposed to work as a basis for communication between a group of agents or between agents and humans. Ontologies are defined as “a shared understanding of some domain of interest” [25]. They also have a predefined structure with an inherent meaning. Their structure consists of a taxonomy, that defines the categories of a domain, relations between these categories, and axiomatizations of categories and relations. We discuss in this paper ontologies particularly for the spatial domain, namely applications for indoor environments in the field of assisted living and ambient intelligence [20].

The extent to which aspects of a domain are defined in an ontology depends, for instance, on intended purposes and applications, granularity, or general design criteria. Different methodologies for the design process of ontologies have been discussed [8], focusing on general engineering frameworks. *Ontology modules*, however, have been introduced from a rather logical point of view [24],

arguing for local consistencies, less complexity, and loose couplings. Here, we adapt this approach to allow the specification of thematically distinct ontologies. Each of these ontologies formulates a particular perspective on the domain. This provides a separation of concerns and encapsulation, and ontologies can benefit from this approach, as it reduces complexity and supports clearly structured specifications. We also show that the use of perspectives for categorizing ontologies support a clearer modularization of these ontologies.

As the ontologies representing the different modules refer to the same domain even though by defining different aspects of it, they are highly related with each other. Particular mappings between ontologies can be defined, for instance, by distributed description logics [3],  $\mathcal{E}$ -connections [16], or ontology matching [6]. These methods range from single mappings of two instances from different ontologies to complex formulas that describe relations between several categories from different ontologies. These approaches relate parts of one ontology to parts of another. Thus they can define link relations between modularized ontologies. In contrast, the formulation of all aspects of the spatial domain in one monolithic ontology is not only difficult to specify and maintain but also increases complexity, limits reasoning practicability, and can cause inconsistencies. Furthermore, a conglomerate of all perspectives on space is hardly ever required by a particular application. Hence, small modules that describe only certain aspects of space provide more flexibility and applicability [9].

This paper is structured as follows. We begin with an application scenario of an ambient assisted living environment as a motivating example, followed by an overview of types of information related to space. We subsequently introduce different modularized spatial ontologies for indoor environments. The application that utilizes the spatial ontologies is then presented, followed by conclusions and future work.

## 2 Motivating Application Scenario

Assisted living environments are supposed to provide building automation, such as control of lighting, air conditioning, appliances, doors, access restriction, and user-based profiles (see, e.g., [15]). Information about architectural elements, such as walls, doors, or windows, provide the most basic kind of architectural data about the floor plan. Functional information about certain rooms or regions, such as the kitchen, dining room, or bathroom, are related to structural parts of the architectural elements. These relations are, however, not fixed: for instance, a room (architectural element) can be related to a bedroom (functional) in the floor plan, though it could later be used as a nursery (functional) or study (functional). Devices that support functions for assisted living are also related to parts of the floor plan, e.g., positions of temperature sensors. One of the functions an assisted living apartment can provide is identifying abnormal heat or cold. For this purpose, the average temperature range is defined a priori by the system, e.g., 18–22°C is considered as an average value in living rooms and 15–20°C in storage rooms with seasonal variations, as part of the temperature control unit of the system.

If temperature values are detected outside this predefined ‘normal’ range, the system has to take into account causes or situations that do not need the system’s reaction (e.g., automatic regulation, alarm, user feedback) in some ways. If the temperature is above average in the bathroom, while the user is taking a shower or bath, the temperature can be ignored, at least temporarily. Here, taking a shower or bath are instances of possible user actions, which have to be defined and recognized by the system. Similarly, if the temperature is above average in the kitchen while the user is cooking, i.e., if the stove is in use and the user is in the kitchen, the situation also seems to be normal (as long as the stove is not in use unintentionally). If, however, the temperature is below average, this could be caused by the user airing the room. In this case, windows of the room or adjacent rooms would be open accordingly. Hence the system has to be able to access the functional parts of the apartment and their conditions or states.

If the constraints that indicate a normal situation of the apartment do not hold, it is likely that a problematic or abnormal situation has been detected by the system. For instance, cold temperatures without open windows may indicate a broken window or heater. As a consequence, certain system reactions have to follow. For instance, the system activates further monitoring sensors for its analysis or tries to regulate the temperature. Or finally, it gets into contact with the user, points out the situation, and asks for further direction.

The examples illustrate the different types of information the apartment has to access and define: (1) architectural building elements, such as walls or windows, (2) functional information of room types, such as kitchen or bathroom, and of assistive devices, such as temperature sensors, (3) types of user actions, such as cooking or taking a shower, (4) types of furniture or devices inside the apartment and their conditions, such as whether the stove is in use, and (5) requirements and constraints of the assisted living system, such as temperature regulations. We claim that these different types of information lead to thematically distinct modules that can be specified as different but related ontologies. Their interactions determine the system’s characteristics and the way it identifies potential abnormal situations implemented as ontological query answering in order to monitor the situation in concrete contexts.

### 3 Perspectives on Space

The thematically distinct modules not only describe particular types of information about the environment. Each module also describes the domain from a certain perspective on a more general level, i.e., the way an ontology describes space not only depends on its selected thematic aspects but also on its selected perspective. These perspectives vary according to the different types of spatial information they describe. They can mainly be divided into four types. First, an ontology may describe space on an abstract or general (and re-usable) level, such as basic locations as specified in *foundational* ontologies. Second, *terminological* aspects specify particular characteristics of space, e.g., for a specific application. Third, an ontology may analyze space from a formal perspective, e.g., by abstract formalizations as in *spatial calculi*. Finally, space can be defined from the

perspective of *multi-modal semantics*, for instance, spatial natural language or gestures. The four different groups categorize possible perspectives from where space can be described:

*Foundational Ontologies.* Ontologies of this group define space often as physical locations of physical entities. The foundational ontology DOLCE [18], for instance, defines these locations as a physical quality of endurants. The spatial location can be described by a physical region, that can, for example, be axiomatized by conceptual spaces. Other general purpose ontologies, such as Cyc [17] and SUMO [19], formalize a variety of spatial theories. The base ontology of SUMO, for example, defines several general spatial categories. One of them is *SpatialRelation*, which represents spatial relations based on mereology and topology. Subcategories of this type include *WhereFn*, a function that maps an object to its position in time and space. [26] introduces a formal theory of space that analyzes locations of entities according to their mereotopological relations between the entities and locations.

Ontologies that describe spatial information from a foundational perspective mainly provide guidelines for the further development of ontologies. Approaches for design decisions of (spatial) ontologies are introduced, for instance, by [28,18]. If different ontologies re-use the same foundational ontology for their top categories, mappings between them can be defined more easily.

*Terminological Ontologies.* Ontologies of this group are often developed for specific purposes or applications. They specify and axiomatize the domain in more detail and less general than the previous group. In the context of geographic information systems, [10] provide a geographic ontology with topological spatial information. A similar approach is provided by [14] with a focus on topographic and environmental information, i.e., hydrology, administrative geography, buildings and places. In the context of visual recognition systems, for instance, [23] have developed a room ontology to recognize indoor scenes.

*Spatial Calculi.* Ontologies of this group define formal calculi by specifying space in an axiomatic and rather abstract way. They do not define terminological aspects of space but abstract entities, such as points, lines, or polygons. An overview of different spatial calculi is given by [5]. A calculus can be reformulated as an ontology [11]. These ontologies specify space, for example, according to region, orientation, shape, distance, origin, or property-specific criteria. In particular, spatial calculi provide composition tables in order to calculate combinations of relations. As spatial calculi are often highly axiomatized, their specification in an ontology is often not directly accessible. Particular reasoners for certain spatial calculi are, however, available [11].

*Multi-modal Semantics.* Ontologies of this group characterize space from a certain (multi-)modal perspective. An example of such an ontology is the formalization of space from a linguistic perspective motivated by the way natural language categorizes the domain. They are used as an interface for natural language

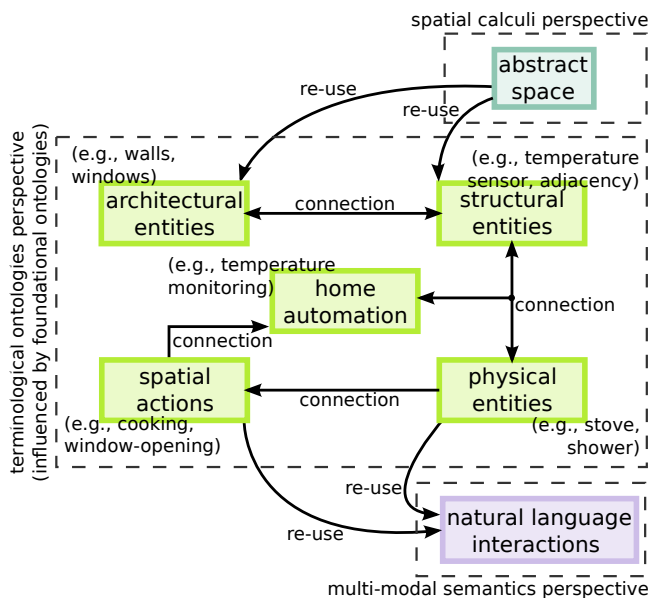
interaction, and they act as an intermediate between the terminological representation and lexicogrammatical information. An example of such a linguistic ontology is presented by [2]. Different ontologies of linguistic semantics may also be specified, for instance, to provide semantics for different languages.

In summary, a perspective determines how the domain is described and which aspects have to be taken into account. Applications can then select those ontologies that satisfy their requirements. In our environment for assisted living, we put this approach into practice by defining and re-using different ontologies for space. The domain is characterized specifically for indoor environments in the context of assisted living. Furthermore, we use spatial calculi and terminological information for ambient assisted living information. Terminological ontologies are categorized according to their thematic aspects and developed on the basis of foundational ontologies that provide design criteria and guidelines. Linguistic semantics can then be used to support natural language interaction.

## 4 Spatial Ontologies and Their Connections for Indoor Environments

The motivating example of an ambient assisted living environment above shows that the domain can be described not only from different perspectives (foundational, terminological, calculus, multi-modal) but also focusing on particular aspects, which can be formulated in ontological modules. A modularized ontology framework for the domain of ambient assisted living needs to define at least the following different aspects: architectural and structural features of buildings, spatial actions and change, indoor objects (physical entities) of the environment, home automation, communication about space (natural language interactions), and abstract space (spatial calculi). An overview of these ontologies is illustrated in Figure 1. Ontologies either re-use imported ontologies or they define connections to other ontologies.

As all ontologies define the same environment in a certain way, all of them refer to categories of the same real world. Even though connections can thus be defined between all ontologies, only those that support concrete application tasks need to be selected. Here, we focus on connections necessary for an intelligent building application. The ontologies are split into functional perspectives on the domain. They provide a separation of concerns in order to allow clear grouping of information, information hiding, and encapsulation. The ontologies keep apart different interpretations and meanings of categories and relations, for example, two ontologies may specify the same floor plan but one with a structural and one with a navigational focus. This can also prevent contradicting implications, as we describe in the Physical Entities section below. A basis for communication and exchange is also provided between different user groups; the way an architect, an interior designer, a construction worker, a painter, or a resident talk about walls differs because of their different perspectives represented by distinct ontologies.



**Fig. 1.** Thematically different spatial ontologies with different perspectives on the domain for indoor assistance applications

The ontologies we describe below are formulated in the web ontology language OWL 2<sup>1</sup>. Thus we focus on ontologies as theories formulated in description logic [1], as they are widely used and a common standard for ontology specifications. Moreover, they provide a balance between expressive power and computational complexity in terms of reasoning practicability. Connections between ontologies are formalized with  $\mathcal{E}$ -connections by defining link relations as object properties, re-use relations between ontologies are formalized with ontology imports.

*Abstract Space.* The module for *abstract space* is primarily used in conjunction with the *structural* and *architectural entities* ontologies. It provides axiomatizations of regions with topological information, based on the region connection calculus RCC-8 [21]. This ontology defines topological restrictions and allows region-based reasoning (simplified in DL). While the architectural ontology defines, for instance, *Door*, *Wall*, and *Ceiling* as they are used in the constructional floor plan, the region-based ontology only defines the category *Region*. An instance D of the category *Door* in the architectural module has a corresponding instance R of the category *Region* in this module and can then be a part of different parthood, connection, or disjointness relations. The RCC-based ontology consists of one category and 15 relations resembling the ontology by [10].

<sup>1</sup> <http://www.w3.org/2007/OWL/wiki/Syntax>

*Architectural Entities.* The modularized ontology for *architectural entities* defines space according to aspects for designing and constructing buildings. Such information can also be used to support architectural design [12]. Entities that are used in construction plans are covered by this module. A Door, for instance, is defined as an area that is attached to wall, floor, or ceiling areas. It is defined by its spatial extent (length, height, and width) and an opening radius. Possible occurrences of doors based on their connections to other areas are constrained in this ontology. Categories reflect mostly geometric features here, but also architecture-specific elements, e.g., the distinction between Wall inside the building and CurtainWall on the outer parts of building. In our application, this module closely resembles a data model of construction plans, the Industry Foundation Classes<sup>2</sup>, and basically provides an ontological interface to the actual construction plan. It consists of about 190 categories and 30 relations.

*Structural Entities.* The modularized ontology for *structural entities* defines space reflecting functional information about the building floor plan. It defines rather abstract containment relations between rooms, floors, and buildings. Hence, this module provides spatial containment relations (parthood in the abstract space module). It also reflects navigation-related information of the environment.

This structural representation of indoor environments does not contain walls defined as an area, as in the architectural module. The structural entities module defines containment relations among the different entities. Door, in this module, is defined as a connection between Rooms, Floors, or Buildings. It provides an entrance to the other entities, while attachments of doors to walls, for instance, are not defined here. A Door  $D'$  in this module is related to its corresponding instance  $R$  in the abstract space module and  $D$  in the architectural entities module. While  $D'$  is the entrance to a room,  $D$  is attached to a specific wall area. The ontology consists of about 90 categories and 20 relations.

*Physical Entities.* The modularized ontology for *physical entities* describes all entities that may occur in the domain of indoor environments. The entities can primarily be defined as those entities that are visually perceivable. Examples of categories in this module are furniture, equipment, appliances, and users, as they are essential for ambient environments in residential buildings. This ontology is developed on the basis of DOLCE-Lite [18], and it consists of about 150 categories and 70 relations. A Door is classified as a physical object with a specific spatial location that exists at a certain time. It defines several subcategories, such as SlidingDoor, RevolvingDoor, or SwingDoor.

Even though physical entities are classifiable into such groups, they may provide other non-prototypical functions that can again be physical entities. A Door, for instance, could be used as an InformationBoard if information signs are attached to the door. Physical roles and artifacts, as they are defined in [27], have to be taken into account here as well. Hence, this module defines entities mostly

---

<sup>2</sup> <http://www.iai-tech.org/ifc/IFC2x4/alpha/html/index.htm>



according to their visual but also according to functional characteristics. An application in the domain of intelligent buildings may be equipped with visual sensors connected to an object recognition system for home automation, as, for instance, presented by [23]. The recognition system may then interact only with this and the *spatial actions* module. Recognized objects are classified according to the categories defined in this module.

The different ontologies illustrated in Figure 1 also prevent ‘contradicting’ definitions: Doors in the *structural entities* ontological module are defined as (entrance) connections between rooms, that are consequently externally connected to the connected rooms. Doors in the *physical entity* ontological module, however, are defined on the basis of their physical properties, i.e., size, weight, material, or color. While the structural door needs to be connected with at least one room, the physical door does not. While the physical door can be made of a certain material, the structural door cannot. Hence, the definition of Door in one module contradicts and does not match with the definition of Door in the other module, as both categories cannot be subsumed by each other. Consequently, an instance of Door in one module would not be an instance of Door in the other module. Different definitions are therefore separated into different ontologies, even though they describe the same door in “reality”, and the contradiction is avoided by a link relation between the ontologies. Two instances for the same door entity from two thematically different ontologies are then linked with each other.

*Spatial Actions.* The modularized ontology for *spatial actions* defines possible actions that can be performed in indoor environments. It also covers change in the environment over time. This module defines actions of users, events, and states of entities. It is related to possible interactions between physical entities from the previous module.

An instance of Door can be open, closed, or locked. This status may change because of an event or a user interaction; automatic access restrictions lock and unlock doors automatically, while users leave doors open or closed. Visual sensors recognize changing environmental aspects and can cause further actions, which are given by ontological restrictions in the home automation module. The action module, which defines about 100 categories and 230 relations, is based on the *Descriptions and Situations* ontology [7].

*Home Automation.* The modularized ontology for *home automation* specifies ontological restrictions on entities from *structural entities*, *physical entities*, and *spatial actions*. In contrast to the other ontologies, particular definitions, e.g., user-based access restrictions, depend on the specific building they are applied to. Hence, they tend to be re-defined in different environments. In the home automation module, access restrictions can be specified. For instance, on the basis of user groups this ontology defines who (physical entities) is allowed to enter (spatial actions) which building element (structural entities). The ontology consists of about 14 categories and 5 relations.

*Natural Language Interaction.* In order to interact with the system using natural language, linguistic semantics for space can be used. Categories and relations that are defined by this module are motivated only on the basis of linguistic characteristics. For human-computer interaction, the physical entities and spatial actions ontologies can be used and related to the linguistic categories, cf. [22].

A Door in the linguistic ontology is, for instance, defined as an element that may participate in a certain linguistic context by playing a certain role, e.g., as an *object*, a *mover*, or a *reference object*. A linguistic ontology that specifies this kind of information is the Generalized Upper Model for Space [2], which consists of 278 categories and 112 relations.

In summary, the ontologies described are modularized according to their thematically different semantic description of the domain. These ontologies reflect a foundational, terminological, abstract, or (multi-)modal perspective on space, as illustrated in Figure 1.

## 5 Application: Ambient Assisted Living

The different ontologies have been designed and customized for the indoor environments of ambient assisted living apartments.<sup>3</sup> An example of this kind of *intelligent apartment* is the Bremen Ambient Assisted Living Lab (BAALL) [15], primarily suitable for the elderly and people with physical or cognitive impairments. In [13], we showed how the apartment can be described spatially and how this description can be used for spatial reasoning about regions or orientations. In the following, we demonstrate the formalization of the temperature monitoring function motivated above, which shows how the ontologies and their connections can be used for providing assisted living functions.

The construction data of the floor plan from Figure 1 gets instantiated in the architectural entities ontology: each element (walls, windows, doors, etc.) and its spatial topological relations are defined. For instance, the five temperature sensors with their positions are instantiated and (RCC-based) externally connected to their respective walls. Functional aspects of parts of the apartment, such as the different types of rooms (kitchen, bathroom, bedroom), are instantiated in the structural entities ontology. These instances are related to the construction parts they are composed of by the link connection between both ontologies. The bathroom, for instance, is composed of its four walls, the door and the window, and it contains a temperature sensor. The physical entities module then instantiates furniture, appliances, and equipment that is part of the apartment and contained in certain rooms. The bathroom, for instance, is linked to its counterpart in the structural entities ontology, and it contains entities, such as a shower, a bathtub, a washbowl, etc., in the physical entities ontology. Actions that can be analyzed by the system are, for instance, the use of water, heating, and light, as specified in the spatial actions ontology.

<sup>3</sup> Our ontologies related to the different modules and perspective in Figure 1 are available at

<http://www.informatik.uni-bremen.de/~joana/ontology/SpatialOntologies.html>, preferred viewer: Protege\_4.0.2.

On the basis of this specification, constraints on the temperature sensors can now be formalized in the home automation module. The following constraint (formulated in Manchester Syntax) reflects the ‘normal situation’ of the apartment’s temperature in bathrooms. The description uses aspects from the modules physical entities (pe), structural entities (se), abstract space (rcc), home automation (ha), spatial actions (sa) and their connections (conn). The statement specifies that a Bathroom in the physical entities ontology satisfies the constraint that it is related to a Bathroom in the structural entities ontology. The statement further specifies that the Bathroom (in pe) is either related to a certain function provided by a Room (in se) that spatially contains a temperature sensor with a value of 18°C-22°C, or in the Bathroom (in pe) happens (in the home automation ontology) a Bathing event (in the spatial actions ontology).

```

Class:      pe:Bathroom
SubClassOf: conn:hasStructuralFeature only se:Bathroom,
            (conn:hasStructuralFeature only
             (inv(se:providesFunction) only
              (se:Room and (rcc:inverseProperPartOf
                           some (se:TemperatureSensor and
                                (se:sensorValue some "18C-22C"))))))
            or (inv(ha:happensIn) some sa:Bathing), ...

```

Ontological reasoning with ABox queries analyzes whether instances satisfy these requirements. Based on an instantiated representation of the apartment given a current situation, the result should show no instances that have temperature sensors in bathrooms that detect values outside the given range and without a bathing action. If, however, such instances exist, they are in conflict with the requirements of the home automation and a certain system reaction follows. In general, the apartment measures its condition regularly and changes the ABox according to the conditions. Reasoning over the ABox can then analyze whether the ontological requirements are still satisfied.

The framework of modularized ontologies not only provides this kind of analyses of a current situation in the environment. Its modules can also interact individually with other systems. For instance, the entrance of the apartment can be equipped with a camera. If a person is detected in front of the main door, a facial recognition system is able to identify the person. This recognition system can especially be combined with the physical entities module that categorizes user groups, and the system can instantiate the recognized person automatically according to the module’s specification. Thus access restrictions in the home automation module can be specified using only the physical entities module and without defining a direct connection to the recognition system.

## 6 Conclusions and Future Work

In this paper, we presented the way modularized ontologies can be structured and interact in order to support functions of assisted living environments. The different ontologies describe the domain based on modularization and perspectives.

Newly created ontologies can thus be developed according to this framework and related to other modules by connections. Applications, such as object recognition or natural language dialogue systems, need to use only those ontologies that are relevant for their functions, such as the physical entities and natural language interactions ontology respectively. Hence, the other modules can be hidden from these applications.

Further work in this field will take into account more ‘dynamic’ aspects. So far, the ambient assisted living environment analyzes snapshots of a current situation by ontological reasoning. However, effects and behavior monitoring of an environment can support further assisted living functions, such as predictions. Moreover, we plan to evaluate the usefulness and scalability of the modularized ontologies by applying and extending them in smart office environments. Future work will also include the investigation of rules by analyzing whether the ontological specification needs to be extended with the SWRL.

## Acknowledgments

This work was supported by the Deutsche Forschungsgemeinschaft (DFG), Collaborative Research Center for Spatial Cognition SFB/TR8, project I1-[OntoSpace]. The author would like to thank John A. Bateman and Bernd Krieg-Brückner for fruitful discussions.

## References

1. Baader, F., Calvanese, D., McGuinness, D., Nardi, D., Patel-Schneider, P.: *The Description Logic Handbook*. Cambridge University Press, Cambridge (2003)
2. Bateman, J.A., Hois, J., Ross, R., Tenbrink, T.: *A linguistic ontology of space for natural language processing*. *Artificial Intelligence* (2010) (in Press)
3. Borgida, A., Serafini, L.: Distributed description logics: Assimilating information from peer sources. *Data Semantics*, 153–184 (2003)
4. Brewster, C., O’Hara, K.: Knowledge representation with ontologies: Present challenges - future possibilities. *Int. J. Human Computer Studies* 65(7), 563–568 (2007)
5. Cohn, A.G., Hazarika, S.M.: *Qualitative Spatial Representation and Reasoning: An Overview*. *Fundamenta Informaticae* 43, 2–32 (2001)
6. Euzenat, J., Shvaiko, P.: *Ontology matching*. Springer, Heidelberg (2007)
7. Gangemi, A., Mika, P.: Understanding the semantic web through descriptions and situations. In: Meersman, R., Tari, Z., Schmidt, D.C. (eds.) *CoopIS 2003, DOA 2003, and ODBASE 2003*. LNCS, vol. 2888, pp. 689–706. Springer, Heidelberg (2003)
8. Gómez-Pérez, A., Fernández-López, M., Corcho, O. (eds.): *Ontological Engineering – with examples from the areas of Knowledge Management, e-Commerce and the Semantic Web*. Springer, Heidelberg (2004)
9. Grau, B.C., Horrocks, I., Kazakov, Y., Sattler, U.: A logical framework for modularity of ontologies. In: Veloso, M.M. (ed.) *20th Int. Joint Conference on Artificial Intelligence*, pp. 298–303 (2007)
10. Grütter, R., Scharrenbach, T., Bauer-Messmer, B.: Improving an RCC-derived geospatial approximation by OWL axioms. In: Sheth, A.P., Staab, S., Dean, M., Paolucci, M., Maynard, D., Finin, T.W., Thirunarayan, K. (eds.) *ISWC 2008*. LNCS, vol. 5318, pp. 293–306. Springer, Heidelberg (2008)

11. Haarslev, V., Lutz, C., Möller, R.: Foundations of spatioterminological reasoning with description logics. In: Cohn, A.G., Schubert, L.K., Shapiro, S.C. (eds.) *Principles of Knowledge Representation and Reasoning: 6th Int. Conference*, pp. 112–123. Morgan-Kaufmann Publishers, San Francisco (1998)
12. Hois, J., Bhatt, M., Kutz, O.: Modular ontologies for architectural design. In: 4th Workshop on Formal Ontologies Meet Industry (FOMI 2009). *Frontiers in Artificial Intelligence and Applications*, vol. 198, pp. 66–77. IOS Press, Amsterdam (2009)
13. Hois, J., Dylla, F., Bhatt, M.: Qualitative spatial and terminological reasoning for ambient environments - recent trends and future directions. In: Bhatt, M., Guesgen, H. (eds.) *1st Workshop on Space, Time and Ambient Intelligence (STAmI 2009)*, pp. 32–43. SFB/TR 8 Spatial Cognition Report Series, No. 020-08/2009 (2009)
14. Kovacs, K., Dolbear, C., Goodwin, J.: Spatial concepts and OWL issues in a topographic ontology framework. In: *GIS Conference* (2007)
15. Krieg-Brückner, B., Gersdorf, B., Döhle, M., Schill, K.: Technology for Seniors to Be in the Bremen Ambient Assisted Living Lab. In: 2. Deutscher AAL-Kongress, VDE-Verlag (2009)
16. Kutz, O., Lutz, C., Wolter, F., Zakharyashev, M.:  $\mathcal{E}$ -Connections of Abstract Description Systems. *Artificial Intelligence* 156(1), 1–73 (2004)
17. Lenat, D.B., Guha, R.V.: *Building Large Knowledge-Based Systems: Representation and Inference in the Cyc Project*. Addison-Wesley, Reading (1990)
18. Masolo, C., Borgo, S., Gangemi, A., Guarino, N., Oltramari, A.: *Ontologies library*. WonderWeb Deliverable D18, ISTC-CNR, Padova, Italy (2003)
19. Niles, I., Pease, A.: Towards a standard upper ontology. In: Welty, C., Smith, B. (eds.) *Formal Ontology in Information Systems*, pp. 2–9. ACM Press, New York (2001)
20. Ramos, C., Augusto, J.C., Shapiro, D.: Ambient intelligence: The next step for artificial intelligence. *IEEE Intelligent Systems* 23(2), 15–18 (2008)
21. Randell, D.A., Cui, Z., Cohn, A.G.: A spatial logic based on regions and connection. In: *3rd Int. Conference on Knowledge Representation and Reasoning*, pp. 165–176. Morgan Kaufmann, San Mateo (1992)
22. Ross, R.J.: Tiered models of spatial language interpretation. In: Freksa, C., Newcombe, N.S., Gärdenfors, P., Wölfl, S. (eds.) *Spatial Cognition VI. LNCS (LNAI)*, vol. 5248, pp. 233–249. Springer, Heidelberg (2008)
23. Schill, K., Zetsche, C., Hois, J.: A belief-based architecture for scene analysis: from sensorimotor features to knowledge and ontology. *Fuzzy Sets and Systems* 160(10), 1507–1516 (2009)
24. Stuckenschmidt, H., Klein, M.C.A.: Integrity and change in modular ontologies. In: Gottlob, G., Walsh, T. (eds.) *18th Int. Joint Conference on Artificial Intelligence (IJCAI 2003)*, pp. 900–908. Morgan Kaufmann, San Francisco (2003)
25. Uschold, M., Grüninger, M.: *Ontologies: Principles, methods and applications*. *Knowledge Engineering Review* 11, 93–155 (1996)
26. Varzi, A.C.: Spatial reasoning and ontology: Parts, wholes, and locations. In: Aiello, M., Pratt-Hartmann, I.E., van Benthem, J. (eds.) *Handbook of Spatial Logics*, pp. 945–1038. Springer, Heidelberg (2007)
27. Vieu, L., Borgo, S., Masolo, C.: Artefacts and roles: Modelling strategies in a multiplicative ontology. In: Eschenbach, C., Grüninger, M. (eds.) *Formal Ontology in Information Systems*, pp. 121–134. IOS Press, Amsterdam (2008)
28. Welty, C., Guarino, N.: Supporting ontological analysis of taxonomic relationships. *Data and Knowledge Engineering* 39, 51–74 (2001)

# Towards Scalable Instance Retrieval over Ontologies

Alissa Kaplunova, Ralf Möller, Sebastian Wandelt, and Michael Wessel

Hamburg University of Technology, 21079 Hamburg, Germany  
<http://www.sts.tu-harburg.de>

**Abstract.** In this paper, we consider the problem of query answering over large multimedia ontologies. Traditional reasoning systems may have problems to deal with large amounts of expressive ontological data (terminological as well as assertional data) that usually must be kept in main memory. We propose to overcome this problem with a new so-called *filter and refine paradigm for ontology-based query answering*.

The contribution of this paper is twofold: (1) For both steps, algorithms are presented. (2) We evaluate our approach on real world multimedia ontologies from the BOEMIE project<sup>1</sup>.

## 1 Introduction

Applying semantic web technologies to enable the semantic retrieval of documents is a hot research topic. We believe that rather expressive DLs such as *SHI* are required in order to capture important domain constraints in an ontology (e.g., less expressive DLs may not provide the required expressivity for the modeling problems at hand). Thus, in this paper we focus on the DL *SHI* (extensions to larger OWL fragments will be considered in future research).

In general, ontologies tend to be large, both in numbers of concepts as well as in numbers of individuals. Unfortunately, the data complexity of instance retrieval in *SHI* (and more expressive DLs) is EXPTIME-complete. Thus, from a computational perspective, instance retrieval with large ontologies (containing lots of instances) may be very hard. Although mature DL/ OWL reasoning systems such as RACERPRO exist [HMW07], many reasoning systems for expressive DLs nowadays still work on main memory only. This obviously prevents their usage on very large ontologies, which may contain millions of “facts”, so query answering simply runs out of main memory, or even loading of the whole ontology is already impossible.

Recently, query answering in less expressive DLs received great attention. E.g., the QUONTO system [ACG<sup>+</sup>05] is able to perform query answering on secondary memory by taking advantage of (relational) database technology.

In this paper, we propose a pragmatic method to combine the *high performance and data scalability* achieved by the techniques realized in the QUONTO architecture with the *high expressivity and expressivity scalability* realized by state-of-the-art DL reasoners such as RACERPRO. We propose a new *filter &*

---

<sup>1</sup> <http://www.boemie.org/>

*refine strategy* for expressive ontologies in order to address the *data- and expressivity scalability problem* [MHW06].

This paper is structured as follows. First, the basics of description logics (as far as relevant for this paper) are introduced; i.e., the DLs *SHI* and *DL-Lite*, as well as basic inference problems. Then, we describe the novel approximation algorithm which reformulates *SHI* ontologies as *DL-Lite* ontologies for the filter step. We then apply a novel partitioning algorithm for the refine step and perform a preliminary evaluation of our framework applied to the AEO ontology. Open problems are discussed and provide motivation for future research. Finally comes the conclusion and some discussion of related and future work.

This paper is accompanied by a technical report[?] containing full proofs and additional details.

## 2 Basics and Guiding Example

In the following part we will define mathematical notions, which are relevant for the remaining paper.

*The Description Logic SHI.* We assume the syntax and semantics of the description logic *SHI* (also called  $\mathcal{ALCHI}_{\mathcal{R}^+}$ ) and *DL-Lite<sub>F</sub>* as defined in [BCM<sup>+</sup>07] and [CGL<sup>+</sup>07].

With  $Ind(\mathcal{A})$  we denote the set of individuals occurring in  $\mathcal{A}$ . We say that  $\mathcal{O}$  is *inconsistent*, denoted with  $INC(\mathcal{O})$ , if there exists no model for  $\mathcal{O}$ . We say that  $\mathcal{O}$  is *consistent*, denoted with  $CON(\mathcal{O})$ , if there exists at least one model for  $\mathcal{O}$ . Given an individual  $a$  and an atomic concept  $C$ , we have  $\langle \mathcal{T}, \mathcal{A} \rangle \models a : C$  iff  $INC(\langle \mathcal{T}, \mathcal{A} \cup \{a : \neg C\} \rangle)$ .

By *instance retrieval for concept C*, we obtain all individuals  $a \in Ind(\mathcal{A})$ , s.t. we have  $\langle \mathcal{T}, \mathcal{A} \rangle \models a : C$ . We denote the set of instances for a given concept  $C$  with  $concept\_instances(C, \mathcal{A}, \mathcal{T})$ .

In the following we define some additional notions, which will be used throughout the remaining part of the paper. A  $\exists$ -*constraint* is a concept description of the shape  $\exists R.C$ , s.t.  $C$  is an arbitrary concept description. A  $\forall$ -*constraint* is a concept description of the shape  $\forall R.C$ , s.t.  $C$  is an arbitrary concept description.

The *subsumption hierarchy* (so-called *taxonomy*) of parents and children for each concept name can be obtained by classification. For *SHI* ontologies it is possible to compute the subsumption hierarchy in advance given only the TBox  $\mathcal{T}$ , i.e. without the ABox  $\mathcal{A}$ . This is possible since *SHI* does not allow the use of nominals. With  $\sqsubseteq_{\mathcal{T}}: N_C \times N_C$  we denote the precomputed taxonomy obtained by classification, e.g., we have  $\sqsubseteq_{\mathcal{T}}(C, D)$  iff  $\mathcal{O} \models C \sqsubseteq D$  for atomic concepts  $C$  and  $D$ . The role hierarchy of a *SHI*-ontology can be computed in advance given the TBox  $\mathcal{T}$  only as well. With  $\sqsubseteq_{\mathcal{R}}: N_R \times N_R$  we denote the precomputed role hierarchy, e.g. we have  $(R, S) \in \sqsubseteq_{\mathcal{R}}$  iff  $\mathcal{O} \models R \sqsubseteq S$  for roles  $R$  and  $S$ .

An atomic concept  $D$  is a synonym for a concept description  $C$  if we have  $\mathcal{T} \models C \sqsubseteq D$  and  $\mathcal{T} \models D \sqsubseteq C$ . With  $synonyms(C, \mathcal{T})$  we denote the set of atomic concepts, which are synonyms for concept  $C$  with respect to  $\mathcal{T}$ . With

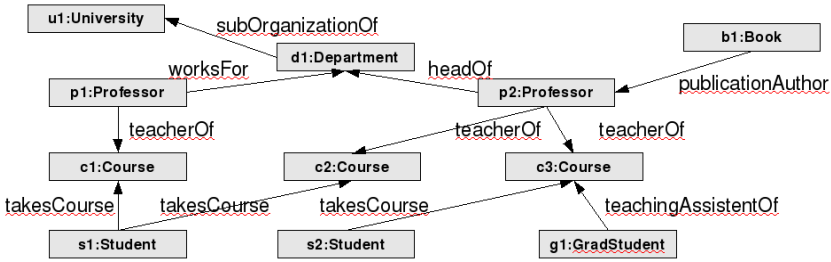


Fig. 1. Guiding Example: ABox  $\mathcal{A}_{EX}$  for ontology  $\mathcal{O}_{EX}$

$\text{parents}(C, T)$  ( $\text{children}(C, T)$ ) we denote the set of atomic concepts which are more general (specific) than a given concept  $C$ .

*Example 1.* In the following we define an example ontology, which is used throughout the remaining part of the paper. The ontology is a simplified version of the LUBM [GPH05]. Let  $\mathcal{O}_{EX} = \langle \mathcal{T}_{EX}, \mathcal{A}_{EX} \rangle$ , s.t.

$$\begin{aligned} \mathcal{T}_{EX} = \{ & Chair \doteq \exists \text{headOf}. Department \sqcap Person, \\ & Professor \sqsubset Faculty, Book \sqsubset Publication, \\ & GraduateStudent \sqsubset Student, Student \doteq Person \sqcap \exists \text{takesCourse}. Course, \\ & \top \sqsubset \forall \text{teacherOf}. Course, \exists \text{teacherOf}. \top \sqsubset Faculty, Faculty \sqsubset Person, \\ & \top \sqsubset \forall \text{publicationAuthor}^-. (Book \sqcup ConferencePaper), \\ & \text{headOf} \sqsubset \text{worksFor}, \text{worksFor} \sqsubset \text{memberOf}, \text{memberOf} \doteq \text{member}^- \} \end{aligned}$$

The ABox  $\mathcal{A}_{EX}$  is shown in Fig. 1. Please note that individual  $p2$  is an non-obvious instance (i.e. we need to perform reasoning) of concept  $Chair$ , since  $p2$  has an outgoing *headOf*-edge to a  $Department$  and every  $Professor$  is necessarily a  $Person$ .

### 3 Terminological Approximation - The Filter Step

*Definition of Approximation.* Let us start with some basic definition. First we define the notion of an approximation of a TBox  $\mathcal{T}$ :

**Definition 1.** For two TBoxes  $\mathcal{T}_1$  and  $\mathcal{T}_2$ ,  $\mathcal{T}_2 \models \mathcal{T}_1$  iff all models of  $\mathcal{T}_2$  are also models of  $\mathcal{T}_1$ .

**Definition 2.** Let  $\mathcal{T}_1$  be a TBox in some DL  $\mathcal{DL}$ . A  $\mathcal{T}_2$  is called an approximation of  $\mathcal{T}_1$  iff a)  $\mathcal{T}_2$  is a  $\mathcal{DL}'$  TBox, with  $\mathcal{DL}' \subseteq \mathcal{DL}$ , and b)  $\mathcal{T}_2 \models \mathcal{T}_1$  holds<sup>2</sup>.

TBox entailment is decidable if  $\mathcal{DL}$  is decidable, since  $\mathcal{T}_2 \models \mathcal{T}_1$  iff for all  $C \sqsubset D \in \mathcal{T}_1$ ,  $C \sqcap \neg D$  is unsatisfiable w.r.t.  $\mathcal{T}_2$ . Note that this is well-defined, since we assume  $\mathcal{DL}' \subseteq \mathcal{DL}$ .

After all, our intention for this definition is that instance retrieval over  $\mathcal{A}$  w.r.t.  $\mathcal{T}_2$  shall be complete, but possibly unsound compared with instance retrieval w.r.t.  $\mathcal{A}$  and  $\mathcal{T}_1$ :

<sup>2</sup> We are discussing the case where  $\mathcal{DL} = SHI$ , and  $\mathcal{DL}' = DL\text{-Lite}$ .



**Proposition 1.** *Let  $\mathcal{A}$  be an ABox which contains only atomic concept assertions, i.e., for all  $i : C \in \mathcal{A}$ ,  $C$  is an atomic concept:  $C \in N_{CN}$ . Let  $D$  be an atomic query concept, the concept whose instances shall be retrieved. Let  $\mathcal{T}_2$  be an approximation of  $\mathcal{T}_1$ . Then, the following holds:  $\text{concept\_instances}(D, \mathcal{A}, \mathcal{T}_1) \subseteq \text{concept\_instances}(D, \mathcal{A}, \mathcal{T}_2)$ .*

*How to Compute Approximations.* Having given these definitions, the question arises, how to actually compute an approximation of  $\mathcal{T}$ . The idea of the *approximation algorithm* is quite simple. W.l.o.g. we assume that a TBox  $\mathcal{T}$  contains only implication axioms (axioms of the form  $C \dot{\sqsubseteq} D$ ; an axiom  $C \dot{\equiv} D$  is transformed into two axioms  $C \dot{\sqsubseteq} D, D \dot{\sqsubseteq} C$ ). Please note that *SHI* admits role inclusion axioms (for roles  $R, S$ )  $R \dot{\sqsubseteq} S$ , which are valid in *DL-Lite* as well. Regarding transitive roles, which are not allowed in *DL-Lite*, the following well-known “trick” from the modal logic realm can be applied:

**Definition 3.** *Let  $R$  be a transitive role in  $\mathcal{T}^3$ . Let  $N_{CN}(\mathcal{T})$  denote the set of concept names appearing in  $\mathcal{T}$ , and  $N_T(\mathcal{T})$  the set of transitively closed roles in  $\mathcal{T}$ . The  $K_4$  closure of  $\mathcal{T}$ ,  $\mathcal{T}^{K_4}$ , is defined as follows:*

$$\mathcal{T}^{K_4} =_{def} \mathcal{T} \cup \{ \exists R. \exists R. C \dot{\sqsubseteq} \exists R. C, \quad \forall R. C \dot{\sqsubseteq} \forall R. \forall R. C \mid C \in N_{CN}(\mathcal{T}), R \in N_T(\mathcal{T}) \}.$$

Moreover, we assume that  $R$  is an ordinary role in  $\mathcal{T}^{K_4} =_{def}$  (not a transitively closed one).

**Proposition 2.** *Let  $D$  be an atomic query concept, and  $\mathcal{A}$  an ABox in which all concept assertions refer to atomic concepts only. Then,  $\text{concept\_instances}(D, \mathcal{A}, \mathcal{T}) = \text{concept\_instances}(D, \mathcal{A}, \mathcal{T}^{K_4})$ .*

We assume a corresponding function `get_K4_closure` which computes the  $\mathcal{T}^{K_4}$  for a given  $\mathcal{T}$ . Please note that this proposition does not hold for arbitrary ABoxes and query concepts  $D$  (only for ABox containing atomic concept assertions, and atomic instance retrieval concepts).

Another preprocessing step is applied to remove nested occurrences of (sub) concepts of the form  $\exists R.C$  and  $\forall R.C$  from the axioms, so they can be better approximated to *DL-Lite* axioms. Thus, for each axiom  $C \dot{\sqsubseteq} D$ , and for each subconcept  $E$  in  $-C \sqcup D$  with  $E = \exists R.F$  or  $E = \forall R.F$ , and  $F \notin N_{CN}$ , we replace  $E$  with a new atomic concept  $C_E$  and add  $\{C_E \dot{\sqsubseteq} E, E \dot{\sqsubseteq} C_E\}$  to  $\mathcal{T}$ . This process continues, until  $\mathcal{T}$  no longer contains such axioms (note that  $E$  itself might still contain such subconcepts as well). For example,  $\{C \dot{\sqsubseteq} D \sqcap \exists R.(E \sqcap F)\}$  is rewritten into  $\{C \dot{\sqsubseteq} D \sqcap C_{\exists R.(E \sqcap F)}, C_{\exists R.(E \sqcap F)} \dot{\sqsubseteq} E \sqcap F, E \sqcap F \dot{\sqsubseteq} C_{\exists R.(E \sqcap F)}\}$ . Consequently, we assume a function `flatten_tbox` which applies this transformation to a TBox  $\mathcal{T}$ . Each model of `flatten_tbox`( $\mathcal{T}$ ) is trivially also a model of  $\mathcal{T}$ , and vice versa, each model of  $\mathcal{T}$  can uniquely be extended to a model of `flatten_tbox`( $\mathcal{T}$ ) (only the new atomic concepts must be interpreted correctly so that their axioms become satisfied).

For an *SHI* TBox  $\mathcal{T}$  we can now compute an approximated  $\mathcal{T}'$  by approximating each axiom. So,  $C \dot{\sqsubseteq} D \in \mathcal{T}$  is replaced by a logically stronger axiom

<sup>3</sup> *DL-Lite* does not offer transitive roles.

$C' \sqsubseteq D'$ ,  $\{C' \sqsubseteq D'\} \models \{C \sqsubseteq D\}$ , which is a *DL-Lite* axiom. The algorithm is best understood as a non-deterministic algorithm which works as follows (the actual deterministic implementation is described briefly below):

```

Function approximate( $T$ )
Parameter:  $\mathcal{SHI}$  TBox  $T$ 
   $T := \text{flatten\_tbox}(\text{get\_K4\_closure}(T))$ 
   $T' := \{C \sqsubseteq D \mid T \models C \sqsubseteq D, C, D \in N_{CN}\}$ 
  while  $T \neq \emptyset$ 
     $\text{axiom} := \text{select\_axiom}(T)$ 
     $T' := T' \cup \text{approximate\_axiom}(\text{axiom}, T')$ 
     $T := T \setminus \{\text{axiom}\}$ 
  end while
  return  $T'$ 

```

The algorithm first syntactically transforms the input TBox  $T$  as explained. Although `flatten_tbox` introduces new atomic concepts, no additional “K4” axioms need to be introduced for them by `get_K4_closure`. Then, the taxonomy of  $T$  is made explicit by adding corresponding axioms to  $T'$ ; these axioms are *DL-Lite* axioms. The reason for this addition to  $T'$  is that the taxonomy of  $T$  shall be available for `approximate_axiom` (see below). Both `select_axiom` and `approximate_axiom` are non-deterministic as well. Given an axiom  $C \sqsubseteq D$ , the basic idea of `approximate_axiom` is to *generalize* the left-hand side  $C$  to  $C'$ , and to *specialize* the right-hand side  $D$  to  $D'$ . This ensures that the approximated axiom is stronger than the original axiom, since  $C' \sqsubseteq D' \models C \sqsubseteq D$  iff  $\neg C' \sqcup D' \models \neg C \sqcup D$  iff  $(\neg C' \sqcup D') \sqcap \neg(\neg C \sqcup D)$  is unsatisfiable iff  $(\neg C' \sqcup D') \sqcap C \sqcap \neg D$  is unsatisfiable iff both  $\neg C' \sqcap C \sqcap \neg D$  and  $D' \sqcap C \sqcap \neg D$  are unsatisfiable. Then, either  $C \sqsubseteq D$  (so this is a tautology, and thus the trivial case), or  $C' \sqsubseteq C'$  (then  $C \sqcap \neg C'$  is unsatisfiable) and  $D' \sqsubseteq D$ , (so  $D' \sqcap \neg D$  is unsatisfiable). In principle, it is of course also sufficient to find equivalent  $C'$ ,  $D'$  in *DL-Lite*. The concepts  $C'$  and  $D'$  are called *possible rewritings* of  $C$  resp.  $D$ , and  $C' \sqsubseteq D'$  is called a *possible rewriting* of  $C \sqsubseteq D$  in the following, or also a *candidate rewriting*.

For example, the axiom  $C \sqsubseteq D \sqcup E$  can be rewritten to  $C \sqsubseteq D$ , or to  $C \sqsubseteq E$  (assuming that  $C, D, E \in N_{CN}$ ). Moreover,  $C \sqsubseteq D \sqcup E$  can also be written as  $\neg D \sqsubseteq \neg C \sqcup E$ ,  $\neg E \sqsubseteq \neg C \sqcup D$ , or even  $\neg D \sqcap C \sqsubseteq E$ , and so on, yielding additional rewriting possibilities. Thus, re-arranging the left-hand sides of the axioms maximizes the number of rewriting possibilities. Even though these axioms are still equivalent to the original one, after rewriting into *DL-Lite* they no longer are. Perhaps for some reordering, no better approximations than  $\top \sqsubseteq \perp$  can be found. It is thus even more important to maximize the number of possible approximations in order to avoid bad approximations which are *too strong* (rendering the whole TBox unsatisfiable).

The `approximate_axiom` function considers the input axiom  $C \sqsubseteq D$  as a disjunction  $\neg C \sqcup D$  which, in a first step, is brought into *disjunctive normal form (DNF)*. A concept is in DNF if it is in *negation normal form (NNF)*, and does not contain any (sub)concepts of the form  $D \sqcap (E \sqcup F)$ . Using simple boolean algebra, each concept can be brought into DNF. Note that the concepts are even simpler at this

step in the processing chain, because complex qualification concepts have been removed in advance. In the following, we use the set notation for disjuncts of a concept in DNF:  $\text{DNF}(C \sqcap (E \sqcup F)) = (C \sqcap E) \sqcup (C \sqcap F) = \{C \sqcap E, C \sqcap F\}$ . The function `approximate_axiom` non-deterministically chooses a subset of  $\text{DNF}(\neg C \sqcup D)$  as a possible left-hand side of the axiom, and uses the remaining disjuncts as right-hand side. Then, `approx_axiom` calls the non-deterministic functions `generalize` and `specialize`:

**Function** `approximate_axiom(axiom, T')`

**Parameter:** *SHI* axiom  $axiom = C \sqsubseteq D$  and partial approximation  $T'$

if  $T' \models axiom$  then return  $T'$

else if  $axiom$  is a *DL-Lite* axiom then return  $\{axiom\} \cup T'$

else

$concept := \text{DNF}(\neg C \sqcup D)$

$left\_side := \text{some\_subset\_of}(concept)$

$right\_side := concept \setminus left\_side$

$left\_side' := \text{generalize}(\neg left\_side, T')$

$right\_side' := \text{specialize}(right\_side, T')$

if  $left\_side' \neq \emptyset$  and  $right\_side' \neq \emptyset$  then

$axiom' := left\_side' \sqsubseteq right\_side'$

if  $T' \not\models axiom'$  then return  $\{axiom'\} \cup T'$

return  $T'$

Both `specialize` and `generalize` first bring their argument concepts in DNF, and then specialize or generalize using a set of *non-deterministic rewriting rules* which are guided by the structure of the concept. The rules are applied exhaustively to the concept  $C$  until no more rule is applicable.

The rules make use of the helper function `syns_or_parents` which returns a non-empty result for *non-atomic concepts only*:

$$\text{syns\_or\_parents}(C, T') =_{def} \begin{cases} \text{synonyms}(C, T') & \text{if } \text{synonyms}(C, T') \neq \emptyset, C \notin N_{CN} \\ \text{parents}(C, T') & \text{if } \text{synonyms}(C, T') = \emptyset, C \notin N_{CN} \\ \emptyset & \text{otherwise} \end{cases}$$

Note that  $T'$  is only partially available, but already contains the taxonomy axioms derived from  $T$  (see `approximate`). Note that  $C \in \text{synonyms}(C, T, T')$  for all  $C \in N_{CN}$ .

The function `generalize` uses the following non-deterministic *generalization rules*;  $C \rightarrow_G C'$  means that  $C$  is generalized to  $C'$ :

- $C \rightarrow_G C'$ , if  $C$  is a valid left-hand side of a *DL-Lite* axiom
- $\exists R.C \rightarrow_G C'$ ,  $C' \in \{\exists R.\top\} \cup \text{syns\_or\_parents}(\exists R.C, T, T')$  (note:  $C \in N_{CN}$ )
- $C \sqcap D \rightarrow_G C'$ ,  $C' \in \{C, D\} \cup \text{syns\_or\_parents}(C \sqcap D, T, T')$
- $C \sqcup D \rightarrow_G C'$ , where  $C' = C_1 \sqcup D_1$ , with  $C \rightarrow_G C_1$ ,  $D \rightarrow_G D_1$ ,  
or  $C' \in \text{syns\_or\_parents}(C \sqcup D, T, T')$
- for all other concepts  $C$ :  $C \rightarrow_G C'$ ,  $C' \in \text{syns\_or\_parents}(C, T, T')$

To give an example, consider `generalize` is applied to  $\exists R.C \sqcup (E \sqcap F)$ . First, the DNF is computed:  $(\exists R.C \sqcap E) \sqcup (\exists R.C \sqcap F)$ . Then, a possible rewriting

is:  $(\exists R.C \sqcap E) \sqcup (\exists R.C \sqcap F) \rightarrow_G \exists R.\top \sqcup F$ , since  $(\exists R.C \sqcap E) \rightarrow_G \exists R.\top$  and  $(\exists R.C \sqcap F) \rightarrow_G F$ . There are many other different rewritings.

Please note that *DL-Lite* does not permit negation or conjunctions on the left-hand sides of axioms; thus, it is impossible to generalize conjunctions by generalizing the arguments analog to the  $\sqcup$ -case. Note that, from this definition, in most cases  $\forall R.C \rightarrow_G \top$  unless `syms_or_parents` finds some parent for  $\forall R.C$  in  $\mathcal{T}'$ . In principle, it is also possible to generalize a disjunction  $C \sqcup D$  to something like  $C \sqcup D \sqcup E$ , for some  $E \in N_{CN}$  (although this will result in a huge search space in the implementation). The rules are designed in such a way to avoid *over-generalization* in order to keep the number of unsound query answers small. That means, more specific rewriting alternatives shall be favored over less specific ones. For example, although  $C \sqcap D \rightarrow_G C \sqcup D$  is conceivable, it doesn't make much sense under this premise, since both  $C \sqcap D \rightarrow_G C$  as well as  $C \sqcap D \rightarrow_G D$  are more specific.

The rules for concept specialization, `specialize`, exploit a similar function `syms_or_children` and follow the principle to avoid *over-specialization*, i.e., more general rewriting alternatives are preferred over more specific ones. In these rules, there is the possibility to rewrite a concept  $C$  to  $\emptyset$ . In case  $C \rightarrow_S \emptyset$  for a conjunct  $C$  in  $C \sqcap D$ , then  $\emptyset$  is considered as  $\top$ . However, in case  $C$  is a disjunct, then  $\emptyset$  is considered as  $\perp$ . So,  $\emptyset$  serves as the neutral element w.r.t. the surrounding operation:

- $C \rightarrow_S C'$ , if  $C$  is a valid right-hand side for a *DL-Lite* axiom
- $\neg C \rightarrow_S \neg C'$ , where  $C \rightarrow_G C'$  (i.e.,  $C$  is generalized),  
or  $C \in \text{syms\_or\_children}(\exists R.C, \mathcal{T}, \mathcal{T}')$ .
- $\exists R.C \rightarrow_S C'$ ,  $C' = \exists R_C.\top$  with  $\mathcal{T}' := \mathcal{T}' \cup \{R_C \dot{\sqsubseteq} R, \exists R_C.\top \dot{\sqsubseteq} C'\}$ ,  
or  $C' = \exists R.\top$  with  $\mathcal{T}' := \mathcal{T}' \cup \{\exists R.\top \dot{\sqsubseteq} C'\}$ ,  
or  $C \in \text{syms\_or\_children}(\exists R.C, \mathcal{T}, \mathcal{T}')$  (note:  $C \in N_{CN}$ )
- $\forall R.C \rightarrow_S \emptyset$ ,  $\mathcal{T}' := \mathcal{T}' \cup \{\exists R.\top \dot{\sqsubseteq} C'\}$ , where  $C \rightarrow_S C'$
- $C \sqcup D \rightarrow_S C'$ ,  $C' \in \{C, D\} \cup \text{syms\_or\_children}(C \sqcup D, \mathcal{T}, \mathcal{T}')$
- $C \sqcap D \rightarrow_G C'$ , where  $C' = C_1 \sqcap D_1$ , with  $C \rightarrow_S C_1$ ,  $D \rightarrow_S D_1$ ,  
or  $C' \in \text{syms\_or\_children}(C \sqcap D, \mathcal{T}, \mathcal{T}')$
- for all other concepts  $C$ :  $C \rightarrow_S C'$ ,  $C' \in \text{syms\_or\_children}(C, \mathcal{T}, \mathcal{T}')$

In principle, it is possible to use  $C \sqcap D \rightarrow_S C \sqcap D \sqcap E$ , for some  $E \in N_{CN}$ , but the same comments as given above (for  $C \sqcup D$ ) apply. Please note that *DL-Lite* does not permit disjunctions on the right-hand sides of axioms. Moreover, `specialize` has a side-effect on  $\mathcal{T}'$ , since it may introduce additional axioms. For example, the  $\exists R.C$ -rule introduces a new range restriction on  $R$  by adding  $\exists R.\top \dot{\sqsubseteq} C'$  to  $\mathcal{T}'$ . So,  $\exists R.C$  is in fact *generalized* to  $\exists R.\top$ ; however, due to the introduced range restriction  $\exists R.\top \dot{\sqsubseteq} C'$  we get  $\exists R.\top \models \exists R.C$ . In combination this is a specialization of  $\exists R.C$ , as required. Another possibility would be to introduce a subrole  $R_C$ ,  $R_C \dot{\sqsubseteq} R$  with range  $C$ , and rewrite  $\exists R.C$  to  $\exists R_C.\top$ , but this would require a modification of the ABox during instance retrieval.

The rule  $\forall R.C \rightarrow_S \emptyset$  deserves an explanation. The idea here is to completely ignore this (sub)concept on the right-hand side, and instead put a new axiom into  $\mathcal{T}'$  (which is modified per side-effect):  $\mathcal{T}' := \mathcal{T}' \cup \{\exists R.\top \dot{\sqsubseteq} C'\}$ . For example, consider the TBox  $\{C \dot{\sqsubseteq} (\forall R.D) \sqcap E\}$ . Since the left-hand side is already

acceptable, only the right-hand side is rewritten:  $(\forall R.D) \sqcap E \rightarrow_S \top \sqcap E$ , since  $\forall R.D \rightarrow_S \emptyset$  and  $E \rightarrow_S E$ . However, also  $\exists R^-. \top \sqsubseteq D$  has been added to  $\mathcal{T}'$ , thus the approximation is  $\mathcal{T}' = \{C \sqsubseteq E, \exists R^-. \top \sqsubseteq D\}$ . It is easy to see that  $\mathcal{T}' \models \mathcal{T}$  holds. In case the input TBox is  $\{C \sqsubseteq (\forall R.D) \sqcup E\}$ , then the following rewriting is possible:  $(\forall R.D) \sqcup E \rightarrow_S \forall R.D \rightarrow_S \emptyset$ . Since `approximate_axiom` will reject axioms with *right\_side'* =  $\emptyset$ , the approximation is simply  $\mathcal{T}' = \{\exists R^-. \top \sqsubseteq D\}$ . Another possibility is of course  $\mathcal{T}' = \{C \sqsubseteq E\}$ , according to the  $\sqcup$ -rule.

**Proposition 3.** *Let  $\mathcal{T}' = \text{approximate}(\mathcal{T})$  for a SHI TBox  $\mathcal{T}$ . Then,  $\mathcal{T}'$  is a DL-Lite approximation of  $\mathcal{T}$ .*

*An Example Approximation.* If `approximate` is applied to the example TBox  $\mathcal{T}$ , then after the preprocessing only the following non-DL-Lite axioms remain which thus have to be approximated:

$$\{ \top \sqsubseteq \forall \text{publication.Author}^-. (\text{book} \sqcup \text{conferencePaper}), \top \sqsubseteq \forall \text{teacherOf.course}, \\ \text{person} \sqcap \exists \text{takesCourse.course} \sqsubseteq \text{student}, \text{student} \sqsubseteq \exists \text{takesCourse.course}, \\ \text{person} \sqcap \exists \text{headOf.department} \sqsubseteq \text{chair}, \text{chair} \sqsubseteq \exists \text{headOf.department} \}$$

One possible DL-Lite approximation  $\mathcal{T}'$  is:

$$\{ \text{chair} \sqsubseteq \exists \text{headOf} \top, \exists \text{headOf} \top \sqsubseteq \text{chair}, \text{chair} \sqsubseteq \text{person}, \\ \text{professor} \sqsubseteq \text{faculty}, \text{book} \sqsubseteq \text{publication}, \text{faculty} \sqsubseteq \text{person}, \\ \text{graduateStudent} \sqsubseteq \text{student}, \text{student} \sqsubseteq \exists \text{takesCourse} \top, \text{student} \sqsubseteq \text{person}, \\ \exists \text{teacherOf}^-. \top \sqsubseteq \text{course}, \exists \text{takesCourse} \top \sqsubseteq \text{student}, \exists \text{teacherOf} \top \sqsubseteq \text{faculty}, \\ \exists \text{publication.Author} \top \sqsubseteq \text{book}, \\ \exists \text{headOf}^-. \top \sqsubseteq \text{department}, \exists \text{takesCourse}^-. \top \sqsubseteq \text{course} \}$$

If this  $\mathcal{T}'$  is used for retrieval on the example ABox, then no unsound answers to atomic instance retrieval queries are delivered. Thus, the approximation is *perfect* for every atomic concept of the ABox. Of course, this depends on the ABox. Note that *p2* is a *chair* instance, as required, since  $\exists \text{headOf} \top \sqsubseteq \text{chair} \in \mathcal{T}'$ .

However, there also exist imperfect approximations. On the one hand, there are many  $\mathcal{T}'$ 's containing incoherent concepts, or even unsatisfiable  $\mathcal{T}'$ 's. In the latter case, the ABox is definitely inconsistent (unsatisfiable), and in the former case, inconsistency of the ABox is likely (if the ABox contains instances of incoherent concepts). From a logical perspective, an instance retrieval query performed on an inconsistent ABox returns the set of *all* ABox individuals (since, from an inconsistent theory, everything follows). So, such a query answer is still complete.<sup>4</sup> On the other hand, an example for an unsound approximation is given by  $\mathcal{T}'$ , if  $\top \sqsubseteq \forall \text{publication.Author}^-. (\text{book} \sqcup \text{conferencePaper})$  is approximated to  $\exists \text{publication.Author} \top \sqsubseteq \text{conferencePaper}$  instead of  $\exists \text{publication.Author} \top \sqsubseteq \text{book}$ . Then, *b1* will be a false answer to the *conferencePaper* instance retrieval query.

Even for the simple example TBox we get 757 coherent approximations (there are a few hundred thousand consistent approximations containing incoherent concepts); w.r.t. retrieval, there is one perfect approximation (see above), and

<sup>4</sup> Of course, a DL reasoner typically does not permit ABox retrieval on inconsistent ABoxes.

the worst coherent approximation has an average failure of 2.5 individuals which means that an atomic instance retrieval query, in the average, returns 2.5 false query answers.

*An Implementation of the Approximation Algorithm.* We have eliminated the non-determinism in the `approximate` algorithm by implementing it in a depth-first (backtracking) search algorithm. Thus, for a given *axiom*, `approximate_axiom(axiom)` returns a set of *candidate axioms*, representing possible approximations of *axiom*. Each axiom thus represents a state in the search space, whose branching factor is given by the number of its candidate approximation axioms.

In principle, the number of possible approximations is truly astronomic for larger TBoxes. Consider a TBox with 500 axioms to approximate, in which each axiom can be approximated in three different ways – the number of atoms in the universe is  $10^{80} \approx 3^{167.6722}$  and thus tiny compared to the  $3^{500}$  nodes in this search space. Thus, clever heuristics are needed to guide the search. Since, in principle, one is only interested in coherent approximations (containing only one incoherent concept name,  $\perp$ ), it is a good idea to prune a path in the search tree as soon as more than one incoherent concept is discovered in the partial  $\mathcal{T}'$ . Of course, this requires a TBox coherence check by the DL reasoner (RACERPRO) at each step. This would be a good use case for *incremental* reasoning. In order to filter out candidate axioms which are too specific, RACERPRO is used as well.

It may not be possible to compute a coherent approximation at all. In this case, an incoherent TBox is wanted which at least leaves the ABox satisfiable (but even this may be impossible), or contains only a minimal number of incoherent concepts.

Sometimes it is possible to compute more than one approximation. Even if each computed approximation is unsound for retrieval on an actual ABox, it is good to have a multitude of approximations available, since their retrieval results can be intersected. Even if no – w.r.t. an actual ABox – perfect approximation is among the computed approximations, this intersected answer set may be perfect for some concept  $C$  on this ABox.

## 4 Island-Based Instance Retrieval – The Refine Step

In the following section we discuss how to post-filter individuals, which were obtained by the Filter Step before. The original algorithm was proposed in [WM08] for the DL *ALCHI*. This section is only intended as an overview of the refine step. Detailed explanations and proofs are omitted in this paper.

The idea for the refine step is that only a subset of role and concept assertions is necessary/used to perform instance checking for a particular given individual  $a$  and a given concept  $C$ . The approach undertaken here is to identify role assertions which can be used during the application of a tableau algorithm for instance checking (note that  $\langle \mathcal{T}, \mathcal{A} \rangle \models^? a : C$  can be reduced to checking whether  $\langle \mathcal{T}, \mathcal{A} \cup \{a : \neg C\} \rangle$  is unsatisfiable via a tableau algorithm).

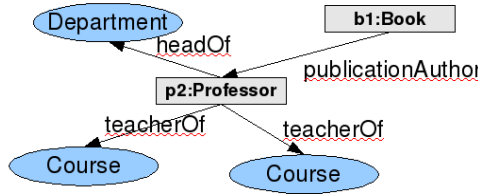


Fig. 2. Example island for individual  $b1$  in  $\mathcal{O}_{EX}$

First, we transform the ontology into some kind of normal form, called *shallow normal form*. For the details of the transformation please refer to [WM08]. We only provide an example for  $\mathcal{T}_{EX}$  from Example 1 in Shallow Normal Form. The TBox  $\mathcal{T}_{EX}$  in SNF is as follows:  $Shallow(\mathcal{T}_{EX}) =$

$$\{ \neg Chair \sqcup \exists headOf.Department, \neg Chair \sqcup Person, \forall headOf.\neg Department \sqcup \neg Person \sqcup Chair, \neg Professor \sqcup Faculty, \neg Book \sqcup Publication, \neg GraduateStudent \sqcup Student, \neg Student \sqcup Person, \neg Student \sqcup \exists takesCourse.Course, \neg Person \sqcup \forall takesCourse.\neg Course \sqcup Student, \forall teacherOf.Course, \forall teacherOf.\perp \sqcup Faculty, \neg Faculty \sqcup Person, \forall publicationAuthor^{\neg}.(Book \sqcup ConferencePaper) \}$$

Given the shallow normal form, we use a so-called  $\forall$ -info structure for an ontology  $\mathcal{O}$  to determine which concepts are (worst-case) propagated over role assertions in an ABox. This helps us to define a notion of separability. The following definition of  $\mathcal{O}$ -separability is used to determine the importance of role assertions in a given ABox. Informally speaking, the idea is that  $\mathcal{O}$ -separable assertions will never be used to propagate “complex and new information” (see below) via role assertions.

**Definition 4.** Given an ontology  $\mathcal{O} = \langle \mathcal{T}, \mathcal{A} \rangle$ , a role assertion  $R(a, b)$  is called  $\mathcal{O}$ -separable, if we have  $INC(\mathcal{O})$  iff  $INC(\langle \mathcal{T}, \mathcal{A}_2 \rangle)$ , where

$$\mathcal{A}_2 = \mathcal{A} \setminus \{R(a, b)\} \cup \{R(a, i_1), R(i_2, b)\} \cup \{i_1 : C | b : C \in \mathcal{A}\} \cup \{i_2 : C | a : C \in \mathcal{A}\},$$

s.t.  $i_1$  and  $i_2$  are fresh individual names.

The extraction of islands for instance checking in ontology  $\mathcal{O}$ , given an individual  $a$ , is now straightforward. From an individual  $a$  one just follows each non- $\mathcal{O}$ -separable role assertion in the original ABox, until at most  $\mathcal{O}$ -separable role assertions are left. For the details of this algorithm please refer to [WM08]. In Figure 2 we show the island computed for individual  $b1$ . Please recall that  $b1$  potentially was a false answer to the *conferencePaper* instance retrieval query. It is easy to see that the computed island does not entail  $b1 : conferencePaper$  and thus  $b1$  can be eliminated from the set of candidates. In Figure 2 it is also easy to see that  $p2$  can be verified to be an instance of *Chair*, by only taking into account the corresponding island, since the *headOf*-edge was preserved during the computation of the island.

*Extension from DL  $\mathcal{ALCHT}$  to DL  $\mathcal{SHI}$ .* Transitive roles can be easily read off from the TBox by additionally taking into account the role hierarchy. Then, whenever we want to compute the island for an individual w.r.t. DL  $\mathcal{SHI}$ , then we have to additionally “follow” all transitive role assertions. This proposal for the extension to DL  $\mathcal{SHI}$  is quite straight-forward and we do not prove it here.

## 5 Preliminary Evaluation

We have performed an initial evaluation of our algorithms on a version of the AEO ontology of the BOEMIE project. Using RACERPRO, we have transformed this OWL ontology into a DL ontology (= TBox, ABox). The utilized TBox DL is  $\mathcal{ALCHf}$ . It contains 1061 axioms which are already in *DL-Lite*, and 499 axioms which have to be approximated to *DL-Lite*. AEO also contains some so-called number restrictions, which we simply approximate to functional roles in *DL-Lite<sub>F</sub>* (since only  $\leq_1 R$  concepts appear).

The ABox of the AEO version we used is rather small – it only contains 138 individuals (266 concept assertions plus 70 role assertions = 336 assertions). We have chosen this ABox since some interesting reasoning is required in order to retrieve the instances of the concept *HighJump* (similar to the *chair* example, but over 2 role fillers).

As illustrated previously, it is very demanding to approximate a TBox with 499 axioms. Unfortunately, we were not successful to compute a coherent approximation of this AEO ontology in reasonable time. Better heuristics are needed here. The reason for this is a massive number of *disjointness axioms*; e.g., axioms of the form  $A \sqsubset \neg B$ ,  $A \sqsubset \neg C$ ,  $\dots$ . Additionally, `get_K4_closure` introduces another 1110 additional axioms.

We have thus simplified AEO substantially by removing all disjointness axioms and ignoring transitivity (so `get_K4_closure` adds no axioms). With this version, a coherent approximation could be computed within 5 minutes. These simplifications do not affect retrieval. In the average, it returns 0,984 false instances for a concept (w.r.t. to the original AEO). The original AEO contains one instance of *HighJump*, and no instances of *SprintCompetition*. The approximated version is perfect for *HighJump*, but delivers 7 wrong *SprintCompetitions*. The *HighJump* instance is in fact also a (false) *SprintCompetition* here. Thus, 7 islands were computed by the partitioning method, ranging in size from 7 to 45 assertions. The average island contains 33.75 assertions. So, in the average, only one tenth of the assertions from the original ABox have to be loaded in order to verify or falsify the candidates for *HighJump* and *SprintCompetition*. Each instance test requires  $\approx 180$  msecs per candidate, thus, after  $\approx 1,440$  seconds the candidate individuals have been refined. Some additional time is needed to compute the islands. Computation of an islands needs milliseconds only (for such small islands).

Since this ABox was rather small, we expanded the ABox artificially by a factor of 500. We thus created an ABox which contained 500 copies of the original ABox, simply by prefixing the individuals with numbers (0 to 499). We then connected



these 500 separated ABox parts using some new artificial role assertions, resulting in a connected ABox, containing 415330 assertions. The ABox still fits into main memory, because otherwise we could not have performed this experiment (the secondary memory-access is not yet realized). The ABox consistency check already needs 3,5 minutes on this ABox now; retrieval requires  $\approx 34$  seconds (for each concepts). As expected, from the approximated version of the TBox, we got 500 *HighJump* instances, and 3500 *SprintCompetitions*. As expected, the newly introduced artificial role assertions connecting the 500 copies have no influence on the size of the islands. Thus, the average islands size is still 33.75; that this is only 0.0008126068 % of the whole ABox. However, now 4000 candidate tests have to be performed, which will require  $\approx 14$  minutes of RACERPRO reasoning time. Additional time for accessing and loading from secondary memory etc. (since also the island partitioning has to work on secondary memory in the future) is taken<sup>5</sup>

## 6 Conclusions, Related and Future Work

Summing up, the evaluation in the previous section has shown that retrieval will require  $\approx 20$  minutes with our framework. This is not too bad compared with the  $\approx 5$  minutes for retrieval on the original AEO (note that the ABox consistency check needs to be performed only once). For a factor of  $\approx 4$ , we have removed the main memory burden. So, this evaluation should be understood as a first preliminary proof of concept of the ideas conveyed in this paper.

Our framework rests on two central assumptions: (1) it must be possible to compute a coherent approximation of the original ontology in *DL-Lite* (or *DL-Lite<sub>F</sub>*), or another less expressive DL, which allows for secondary memory retrieval. As our preliminary evaluation with a real-world multimedia ontology has shown, this may be very hard. Several problems regarding the efficient handling of disjoint axioms and transitive roles still need to be solved, (2) the original ontology must allow for effective partitioning, so that the individual partition do not exceed main memory size. This may not be the case for all ontologies.

## References

- [ACG<sup>+</sup>05] Acciarri, A., Calvanese, D., De Giacomo, G., Lembo, D., Lenzerini, M., Palmieri, M., Rosati, R.: Quonto: Querying ontologies. In: Proc. of the 20th Nat. Conf. on Artificial Intelligence, AAAI 2005 (2005)
- [BCM<sup>+</sup>07] Baader, F., Calvanese, D., McGuinness, D.L., Nardi, D., Patel-Schneider, P.F.: The Description Logic Handbook. Cambridge University Press, New York (2007)
- [CGL<sup>+</sup>07] Calvanese, D., De Giacomo, G., Lembo, D., Lenzerini, M., Rosati, R.: Tractable Reasoning and Efficient Query Answering in Description Logics: The DL-Lite Family (2007)

---

<sup>5</sup> The test files and results can be downloaded from

<http://www.sts.tu-harburg.de/~mi.wessel/download/boemie-experiment.zip>

- [GPH05] Guo, Y., Pan, Z., Heflin, J.: Lubm: A benchmark for owl knowledge base systems. *J. Web Sem.* 3(2-3), 158–182 (2005)
- [HMW07] Haarslev, V., Möller, R., Wessel, M.: *RacerPro User's Guide and Reference Manual Version 1.9.1* (May 2007)
- [MHW06] Möller, R., Haarslev, V., Wessel, M.: On the scalability of description logic instance retrieval. In: Freksa, C., Kohlhase, M. (eds.) 29. Deutsche Jahrestagung für Künstliche Intelligenz. LNCS (LNAI), Springer, Heidelberg (2006)
- [WM08] Wandelt, S., Moeller, R.: Island reasoning for alchi ontologies. In: *Proceedings of the 5th International Conference on Formal Ontology in Information Systems (FOIS 2004)*. IOS Press, Amsterdam (2008)

# MindDigger: Feature Identification and Opinion Association for Chinese Movie Reviews

Lili Zhao and Chungping Li

School of Software, Tsinghua University,  
Beijing, China  
zhaoll07@mails.tsinghua.edu.cn,  
cli@tsinghua.edu.cn

**Abstract.** In this paper, we present a prototype system called MindDigger, which can be used to analyze the opinions in Chinese movie reviews. Different from previous research that employed techniques on product reviews, we focus on Chinese movie reviews, in which opinions are expressed in subtle and varied ways. The system designed in this work aims to extract the opinion expressions and assign them to the corresponding features. The core tasks include feature and opinion extraction, and feature-opinion association. To deal with Chinese effectively, several novel approaches based on syntactic analysis are proposed in this paper. Running results show the performance is satisfactory.

**Keywords:** Feature Identification, Opinion Association, Syntactic Analysis.

## 1 Introduction

With the rapid development of Web2.0 that emphasizes the participation of the users, more and more Websites, such as IMDb (<http://www.imdb.com>) and Amazon (<http://www.amazon.com>), encourage people to post reviews for the interested items. These reviews contain valuable information which can change the behavior of the users. However, the reviews are massive and unstructured, it is hard for people to find or collect useful information they want, which motivates the research of Opinion Mining. In this paper, we focus on Chinese movie reviews. For movie reviews, there have been work in this domain [1, 12, 15, 16]. Most of them are English expression oriented, which make the proposed approaches not very suitable for Chinese. So Chinese movie review mining is a more challenging task.

In this paper, we aim to design a prototype system which is capable of searching movies, identifying features and analyzing opinions from the reviews. To solve the problem, we decompose the task into the following subtasks: 1) identifying features and opinion words in a sentence; 2) matching the opinion with the related feature. we propose several novel approaches, and a prototype system, MindDigger, has been developed.

The rest of this paper is organized as follows: section 2 introduces related work. Section 3 and section 4 describe the proposed approaches. The system architecture will be introduced in section 5. Section 6 will give the experimental results. Finally, we will conclude the paper in Section 7.

## 2 Related Work

Opinion analysis have been studied by many researchers in recent years. There are two main directions in this field, one is in document level and the other is in feature level. Document level opinion mining [1–3, 8, 12, 14, 15] investigates ways to classify the whole review as positive, negative or neutral. While the later is interested in finding products features being commented on and the opinion polarity for each feature. In this paper, we focus on feature level opinion analysis.

There has been work on feature level opinion analysis [4, 5, 7, 9, 16]. As the pioneer work, [16] classified and summarized the movie reviews by extracting high frequent opinion keywords. Feature-opinion pairs were identified by using a dependency grammar graph. [7] proposed a method that uses word attributes, including occurrence frequency, POS and synset in WordNet. First, the product features were extracted. Then, the features were combined with their nearest opinion words, which are from a generated and semantic orientation labeled list containing only adjectives. Finally, a summary was produced by selecting and re-organizing the sentences according to the extracted features. To deal with reviews in a specific format, Liu et al expanded the opinion word list by adding some nouns [10]. [4] further improved Hu’s system by manually adding some rules to handle different kinds of sentence structures. In the Pulse system introduced by [5], a bootstrapping process was used to train a sentiment classifier. The features were extracted by labeling sentence cluster according to their key terms. In [9], they proposed a machine learning approach which is built under the framework of lexicalized HMMs. Their approach naturally integrates multiple important linguistic features into automatic learning.

## 3 Feature Corpus

This section introduces the feature corpus we use in our system. Before starting, we give the description of movie feature: a movie feature is a term that satisfies one of the following relationships: an element of a movie; an attribute of a movie; an cast member of a movie.

When users comment on product features, the words they use converge [6]. We can draw the same conclusion for Chinese movie reviews according to the statistical results after studying many pieces of reviews. Although labeling features manually can identify features accurately, it is a labor intensive and time consuming task. Thus, if the system can extend feature words with synonyms and identify new features automatically it would be great. Based on this purpose, we design a corpus-based approach which has the advantage of providing a handful of seeds, the system will automatically extend the data set with

synonyms. The seed feature set in our method is pre-defined according to the movie metadata of IMDB. IMDB houses a large collection of published movies. The information presented for each movie include screenplay, character design, vision effects, director, producer, story, special effects maker etc. They can provide specific features of a movie which consumers would be interested in. In this paper, they are translated into Chinese to accomplish the task. We extend the feature set according to the synonyms corpus<sup>1</sup>.

There are 77,345 terms in the Synonyms Corpus. All the terms are organized into a tree structure according to the term's semantic, the Synonyms Corpus provides 5-layer encoding, each encoding has 8 bits, which uniquely represents the terms in same line. For example:

- Ba01A02= (substance)(quality)(element)
- b06E09@ (folk)
- Ba01B10‡ (conductor) (semiconductor)

**Table 1.** Tag Specification of Synonyms Corpus

Encoding Bit	1	2	3	4	5	6	7	8
Encoding Sample	D	a	1	5	B	0	2	= ‡@
Encoding Sense	big	middle	small		general class			
Encoding Level	first level	second level	third level	fourth level	fifth level			

The way of encoding is as follows: the upper character represents both the first and the fourth level, the lower character represents the second level, 2-bits decimal number represents both the third level and the fifth level. Specific tags are shown in Table 1. They are in accordance with the order from left to right. There are 3 tags for the eighth bit: "=", "‡", and "@", which represent the meaning of "(equal, synonyms)", "(unequal, similar)", "(independent which means it has no synonyms or similar words)" separately.

According to the encoding characteristic, we design a term similarity calculation method called TSC (the abbreviation for Term Similarity Calculation). The basic idea of TSC is to compare the 8-bit encodings of two terms. Starting from the highest bit, it checks if the two terms are equal or synonymous with each other. If not, it continues to compare the previous 7 bits until gets a similarity value. The detailed calculation method is as follows:

First, it checks whether the two terms are equal, if it is, then set the similarity value to 1; if not, it will use the 8-bit encodings of the two terms to calculate as follows:

1. If the previous 7 bits of the two encodings are equal and the *8th* bits are "=" or "@", the similarity value is set to 1.
2. If the previous 7 bits of the two encodings are equal and the *8th* bits are "‡", the similarity value is set to 0.5.

<sup>1</sup> The synonyms corpus was provided by Harbin Institute Technology, IR Lab.

3. Considering the current  $i$ th bit ( $i$  ranges from 2 to 8), if the previous ( $i - 1$ ) bits of the two encodings are the same and the  $i$ th bits are different, the similarity value equals  $\frac{1}{10-i}$ .

By using the proposed method of term similarity calculation based on Synonyms Corpus, we can cover terms as many as possible.

## 4 Opinion Association

After studying Chinese expressions of many pieces of movie reviews, we find that 86% of the discussed features are in subject-predicate (SBV) structure, verb-object (VOB) structure, or attribute (ATT) structure. To verify this observation, we examined 250 pieces of movie reviews of five kinds of movies. We found that 86% of the discussed features are in subject-predicate structure, verb-object structure, or attribute structure of a sentence. Therefore, we develop a syntactic analysis based approach for opinion association in Chinese movie reviews. By the new method, feature words are extracted based on feature corpus first, and then all the adverbs and adjectives in the feature bearing sentence are extracted as opinion candidates. The reason why we extract adverbs and adjectives is based on the observation in [11], the authors indicate that in Chinese language, 91% adjectives following an adverb are opinions and 81% opinion words are adjective words following adverbs in sentences.

Due to our observation, in a sentence, we can analyze the SBV structure, and VOB structure, and ATT structure to associate opinion with the related feature. In the following we use OASA for abbreviation of opinion association based on syntactic analysis. Algorithm 1 describes our OASA algorithm. Let the notation  $\text{Opinion}(term)$  to denote the opinion information of the term. For example, the opinion information of the term "good" is  $\text{Opinion}(good)=good$ .

To find the opinions on the discussed features, for each feature we first find the structure bearing the feature. Line 3 to line 21 checks if the feature is in SBV structure. If it is, we continue to check if the predicate is an adjective (line 4). If the predicate is an adjective, line 5 to 8 checks the presence of any negation words in front of the predicate and changes the opinion accordingly. If the predicate is a verb, line 12 finds the CMP structure and pass the opinion of the complement to the feature (line 13). Line 15 indicates that there is no CMP structure found. Then we continue to search VOB structure to determine the opinion of feature(line 15 to 17). Line 22 to line 33 checks if the feature is in ATT structure. If the predicate is an adjective, line 24 to 27 checks the presence of any negation words in front of the predicate and changes the opinion accordingly. If the attribute is a noun, line 30 continues to find ATT structure which contains the attribute, then pass the opinion to the feature. Line 34 to line 43 checks if the feature is in VOB structure. According to the description about feature, it is impossible not a verb. Therefore, in this VOB the feature is the object of the VOB structure. With this conclusion, line 35 searches the VOB structure containing the verb. If the object in the later VOB structure is

---

**Algorithm 1.**OASA

---

**Require:** a set of feature bearing sentences  $FEATURE_SENTENCESET$ .**Ensure:** feature-opinions pairs

```

1: for each sentence in  $FEATURE_SENTENCESET$  do
2:   for each feature in the sentence, Initial Opinion(feature)= null do
3:     if the feature is in SBV structure then
4:       if the predicate is an adjective then
5:         if there is a negative word before the predicate then
6:           Opinion(predicate)=negative predicate,
           Opinion(feature) =Opinion(predicate).
7:         else
8:           Opinion(predicate)=predicate, Opinion(feature)=Opinion(predicate).
9:         end if
10:      end if
11:     if the predicate is a verb then
12:       if complement structure which contains the predicate is found then
13:         Opinion(complement)=complement,
         Opinion(feature)=Opinion(complement).
14:       else
15:         find the VOB structure which contains the predicate,
16:         if the object in VOB structure is an adjective then
17:           Opinion(object)=object, Opinion(predicate)=Opinion(object).
18:         end if
19:       end if
20:     end if
21:     else
22:       Opinion(feature)=null
23:     end if
24:     if the feature is in VOB structure then
25:       find the VOB structure which contains the verb,
26:       if the object is an adjective then
27:         if there is a negative word before the adjective then
28:           Opinion(adjective)=negative adjective
           Opinion(verb)=Opinion(adjective), Opinion(feature)=Opinion(verb).
29:         else
30:           Opinion(adjective)=adjective,
           Opinion(verb)=Opinion(adjective), Opinion(feature)=Opinion(verb).
31:         end if
32:       end if
33:     else
34:       Opinion(feature)=null
35:     end if
36:     if the feature is in ATT structure then
37:       if the attribute is an adjective then
38:         if there is a negative word before the predicate then
39:           Opinion(adjective)=negative adjective,
           Opinion(feature)=Opinion(adjective).
40:         else
41:           Opinion(adjective)=adjective, Opinion(feature)=Opinion(adjective).
42:         end if
43:       if the attribute is a noun, find the ATT structure which contains the
         attribute then
44:         Opinion(adjective)=adjective
         Opinion(attribute)=Opinion(adjective),
         Opinion(feature)=Opinion(attribute)
45:       end if
46:     end if
47:     else
48:       Opinion(feature)=null
49:     end if
50:   end for
51: end for

```

---

an adjective, line 36 to 40 checks the presence of any negation words in front of the predicate and changes the opinion accordingly, then pass the opinion to the feature. By applying algorithm OASA, we can associate most of opinions with the related feature in many simple and normal order sentences. But for some sentences, we can not correctly assign the opinion to the discussed feature. For example, "(This animation reflects the excellent film production.)". The result of syntactic analysis is shown as follows:

```
[4](reflect)-[5](MT) [3](animation)-[2](ATT)
[7]-[6](film production)(DE) [3](animation)-[1](this)(ATT)
[4](reflect)-[3](animation)(SBV) [7]-[4](reflect)(DE)
[8](excellent)-[7](ATT) [9] <EOS>-[8](excellent)(HED)
```

This sentence has two features: animation and film production. Obviously, the opinion word "excellent" is assigned to film production. After running Algorithm OASA, we can observe the result in Table 2 that there are no opinion words assigned to the features.

**Table 2.** The Result by Running Algorithm OASA

Feature Bearing Structure	Feature	Opinion
[7]-[6](film production)(DE)	[6](film production)	null
[4](reflect)-[3](animation)(SBV)	[3](animation)	null

Through analysis, we found that this problem is caused of the lack of consideration on the DE structure which holds features. Moreover, in the movie reviews, many feature bearing sentences are expressed in this way. Therefore, we modified algorithm with the consideration on DE structure. If the feature is in DE structure, we find the nearest adjective according to the POS tagging results , then pass the opinion of the adjective to the feature.

The result of Algorithm OASA with consideration with DE structure is shown in Table 3.

**Table 3.** The Result by Running Algorithm OASA

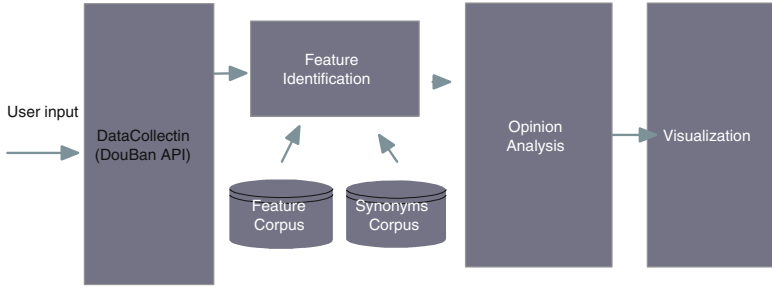
Feature Bearing Structure	Feature	Opinion
[7]-[6](film production) (DE)	[6](film production)	(excellent)
[4](reflect)-[3](animation)(SBV)	[3](animation)	null

## 5 System Description

MindDigger has three major components: Feature Identification, Opinion Analysis and Visualization, which is shown in Fig 1.

Before going to the next section, each review is processed by segmentation, part-of-speech tagging and syntactic analysis software. We use IRLAS2.0 provided by Harbin Institute of Technology to do these tasks.





**Fig. 1.** The System Framework

## 6 Experiment and Analysis

In this section, we evaluate MindDigger to assess its accuracy for feature identification and the performance for opinion association.

### 6.1 Data

We applied MindDigger to a popular website Douban (<http://www.douban.com>). The users can submit movie reviews to the website, with no editing beyond simple filtering. The reviews range in length from a single sentence to hundreds of sentences (one piece of review).

### 6.2 Feature Identification

For the feature identification, we performed the experiments on five genres of movies. They are action, love, comedy, science-fiction, and story respectively. For each genre, we select one typical and familiar movie, then download 250 pieces of reviews. The selected movies are X-Men Origins: Wolverine, Titanic, Ice Age: Dawn of the Dinosaurs, Transformers: Revenge of the Fallen, The Reader. The characteristics of each review data set are given in Table 4. Table 5 gives the experimental results. The performance are measured using the standard evaluation measures of precision:

$$precision = \frac{N(\text{correctly identified features})}{N(\text{all correctly features})}$$

where  $N(*)$  denotes the number of \*. Comparing with the product review mining results reported in [13] and [6], we can find that the precision is much lower than that of product review mining. Since movie reviews are known to be difficult with opinion analysis, especially for the ones written in Chinese. Movie reviews often contain many sentences with feature information about the plot of the movie, although these sentences do not contain any opinion information. Therefore, there are many confusing feature sentences in them, which causes the low precision.

**Table 4.** The Characteristics of the review data

Movie	Number of re-views	Number of Non-repetitive Features	Number of feature sentences(including repetitive features)
X-Men Origins: Wolverine	50	32	203
Titanic	50	27	276
Ice Age	50	20	232
Transformers	50	26	346
The Reader	50	25	263

**Table 5.** Results of feature identification

Features	Number of Feature bearing Sentences	Number of Sentences with Opinion Information	Number of correctly Detected Sentences	Precision
(film)	230	164	96	58.54%
(love)	83	22	15	68.18%
(story)	65	16	8	50%
(subject)	30	12	5	41.67%
(plot)	24	18	6	33.3%
(shot)	21	9	4	44.44%
(screen)	17	9	3	68.18%
(scene)	16	3	1	33%
(action)	13	2	2	100%
(voice)	12	5	4	80%

### 6.3 Opinion Association

In this section, we use manually labeling as the baseline. Therefore, for each sentence with opinion information which are obtained from the feature identification module, we manually associated opinion with the related feature. Then, we ran the algorithm OASA with consideration on DE structures on the sentences. The opinion association results are shown in Table 6. The accuracy produced by algorithm OASA with the consideration on DE structure is much lower than that of product review mining. Through manual inspection, we observed that the algorithm was highly dependent on the syntactic analysis. Moreover, most of the review sentences are non-normal word order and too long which add more difficulty to opinion association problem.

**Table 6.** The result of the opinion association

	OASA	manually labeling	Recall
SBV structure	278	387	71.83%
VOB structure	177	263	75%
ATT structure	365	468	78%

## 7 Conclusion

In this paper, a prototype system called MindDigger is designed for feature identification and opinion association. Several novel approaches based on syntactic analysis are proposed. Specifically, the method naturally integrates syntactic analysis into opinion association. The existing problem are: (1) People like to describe long pieces of movie plot. For example, some people like to describe how the story began, and what was the movement of the movie. This influences the system performance, because the plot description has no opinion information but usually contains some feature words. (2) For some complex sentences, due to the incorrectness from the syntactic analysis, the algorithm OASA can not be carried out very well. This is also left for future work.

## References

1. Chaovalit, P., Zhou, L.: Movie review mining: a comparison between supervised and unsupervised classification approaches. In: HICSS 2005: Proceedings of the 38th Annual Hawaii International Conference on System Sciences (HICSS 2005) - Track 4, Washington, DC, USA, p. 112. IEEE Computer Society, Los Alamitos (2005)
2. Das, S., Chen, M.: Yahoo! for amazon: Extracting market sentiment from stock message boards. In: Proceedings of the 8th Asia Pacific Finance Association Annual Conference (2001)
3. Dave, K., Lawrence, S., Pennock, D.: Mining the peanut gallery: Opinion extraction and semantic classification of product reviews. In: Proceedings of the 12th international conference on World Wide Web, pp. 519–528. ACM, New York (2003)
4. Ding, X., Liu, B., Yu, P.: A holistic lexicon-based approach to opinion mining. In: Proceedings of the international conference on Web search and web data mining, pp. 231–240. ACM, New York (2008)
5. Gamon, M., Aue, A., Corston-Oliver, S., Ringger, E.: Pulse: Mining customer opinions from free text. In: Famili, A.F., Kok, J.N., Peña, J.M., Siebes, A., Feelders, A. (eds.) IDA 2005. LNCS, vol. 3646, pp. 121–132. Springer, Heidelberg (2005)
6. Hu, M., Liu, B.: Mining and summarizing customer reviews. In: Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining, pp. 168–177. ACM, New York (2004)
7. Hu, M., Liu, B.: Opinion extraction and summarization on the web. In: Proceedings Of The National Conference On Artificial Intelligence, vol. 21, p. 1621. AAAI Press/MIT Press, Menlo Park/Cambridge (2006)
8. Java, A.: A framework for modeling influence, opinions and structure in social media. In: AAAI, pp. 1933–1934. AAAI Press, Menlo Park (2007)
9. Jin, W., Ho, H.H., Srihari, R.K.: Opinionminer: a novel machine learning system for web opinion mining and extraction. In: KDD 2009: Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining, pp. 1195–1204. ACM, New York (2009)
10. Liu, B., Hu, M., Cheng, J.: Opinion observer: Analyzing and comparing opinions on the web. In: Proceedings of the 14th international conference on World Wide Web, pp. 342–351. ACM, New York (2005)

11. Liu, H., Yang, H., Li, W., Wei, W., He, J., Du, X.: Cro: a system for online review structurization. In: KDD 2008: Proceeding of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining, pp. 1085–1088. ACM, New York (2008)
12. Pang, B., Lee, L., Vaithyanathan, S.: Thumbs up?: sentiment classification using machine learning techniques. In: Proceedings of the ACL-2002 conference on Empirical methods in natural language processing, vol. 10, pp. 79–86. Association for Computational Linguistics, Morristown (2002)
13. Popescu, A., Etzioni, O.: Extracting product features and opinions from reviews. In: Proceedings of HLT/EMNLP, vol. 5, pp. 339–346. Springer, Heidelberg (2005)
14. Turney, P., et al.: Thumbs up or thumbs down? semantic orientation applied to unsupervised classification of reviews. In: Proceedings of the 40th annual meeting of the Association for Computational Linguistics, pp. 417–424 (2002)
15. Ye, Q., Shi, W., Li, Y.: Sentiment classification for movie reviews in chinese by improved semantic oriented approach. In: Proceedings of the 39th Annual Hawaii International Conference on System Sciences, HICSS 2006, January 2006, vol. 3, p. 53b (2006)
16. Zhuang, L., Jing, F., Zhu, X.: Movie review mining and summarization. In: Proceedings of the 15th ACM international conference on Information and knowledge management, pp. 43–50. ACM, New York (2006)

# The Impact of Latency on Online Classification Learning with Concept Drift

Gary R. Marrs, Ray J. Hickey, and Michaela M. Black

School of Computing and Engineering, University of Ulster, Coleraine,  
County Londonderry, N. Ireland  
marrs-g@email.ulster.ac.uk,  
{mm.black,rj.hickey}@ulster.ac.uk

**Abstract.** Online classification learners operating under concept drift can be subject to latency in examples arriving at the training base. A discussion of latency and the related notion of example filtering leads to the development of an example life cycle for online learning (OLLC). Latency in a data stream is modelled in a new Example Life-cycle Integrated Simulation Environment (ELISE). In a series of experiments, the online learner algorithm CD3 is evaluated under several drift and latency scenarios. Results show that systems subject to large random latencies can, when drift occurs, suffer substantial deterioration in classification rate with slow recovery.

**Keywords:** Online Learning, Classification, Concept Drift, Data stream, Example life-cycle, Latency, ELISE, CD3.

## 1 Introduction

Online learning for classification involves inducing an initial classifier and updating this at intervals from a data stream of time-stamped training examples. Current approaches deploy a variety of machine learning algorithms either individually or in ensembles. They use a variety of means of handling concept and /or population drift and of maintaining a set of valid training examples. See [1], [2] and [3] for an introduction to and review of existing work and underlying issues; the potential for lack of representativeness of examples in a data stream is discussed in [4].

There has been very little discussion in the literature on the time-stamp itself and what it represents. Implicitly it is taken to be the time at which an example becomes available to the learning algorithm for use at some point in the future.

A training example is supposed to be representative of some underlying true rule operating at a particular time. Yet an example may only become available *after* that time. This leads to a time discrepancy or *latency*. Such latencies may impact on learning updates when drift occurs and are the subject of this work.

In section 2, latency is discussed. This leads to an extended model of the life cycle of an example within an online learning system. An experimental test-bed for investigating the impact of latency on learning is described in section 3 and results are presented and analysed in section 4.

## 2 Latency and the Online Learning Life-Cycle

In on-line classification learning, the learning algorithm receives new training examples on a periodic basis and adopts a learning update regime to maintain currency of the classifier. These new examples are time-stamped and placed in a training base. The time-stamps will influence the learner.

The issue as to what the time-stamp of a new training example should be has received little attention in the literature. When a classifier makes a classification at time  $t$ , it is attempting to replicate the classification that would be made by an oracle (possibly a human expert), were such available. This oracle is indexed by time and can change its rules over time, i.e. concept drift. It is the role of the learning regime to try to capture and maintain the oracle's rules over time.

In due course, the true class for this example may be revealed. Consider a credit card fraud prevention system which aims to identify fraud at the time of transaction. Suppose the current classifier accepts the transaction as legitimate. At a later date it may become apparent that the transaction was fraudulent. This information may be returned to the online learning system as a new training example.

When fed to the learner at the next update of learning, what should the time-stamp of this example be? At first sight it might appear that the time-stamp should be that at which the fraud was verified. The training example, however, is to be regarded as a window into the behaviour of the oracle that was current at the time of classification. The time-lapse until verification is seen as purely administrative. Although a decision is made by the company, at the time of verification, to designate the transaction as fraudulent, perhaps after extensive investigation, this decision-making process is distinct from that of the oracle operating at the time of classification. Thus the appropriate time-stamp for the training example, when presented to the learner, is the time of classification.

The time-interval from classification until the verified class for the example is discovered is called its *verification latency*.

In contrast, consider an online medical diagnostic system which receives, from time to time, cases that have been classified by a medical expert. The aim of this system is to emulate the expert. Here the verification latency is zero: there is no gap between the application of the oracle, i.e. the expert, and the appearance of the verified class.

It may transpire in the future that the classification made by the medical expert was wrong, e.g. the patient did not have the disease as diagnosed. Given, however that the purpose of the learner is to emulate the expert, this is not relevant. Were the disease that the patient actually had to be regarded as the verified class and fed back to the learner, this would amount to using nature as the oracle and not the human expert. In this latter situation, the system becomes identical to the fraud system and verification latency should be regarded as the time difference between that when the patient's symptoms were noted and when the correct diagnosis was obtained.

Following verification, there may be a further delay, possibly administrative, before a training example becomes available as a new example to the learner. Such delays can affect both types of systems discussed above.

Clearly latency only becomes an issue should concept drift occur. In a new episode of learning to update the classifier, some or all of the examples received since the last episode will be input to the learner. If drift has occurred from one episode of learning to the next, then some examples will reflect the old oracle and so may be out of date. The extent to which this is the case depends on the magnitude of latency and the time between learning episodes.

Depending on the application, latency can be fixed or random. In a system which predicts rise or fall in a share price from the close of business in a stock market from the end of one day until the end of the next day, verification latency is fixed at one day and there is no administrative delay. In the fraud application, verification latency and administrative delay are likely to be random. With random latency, training examples will become available out of chronological order.

Random latency may be different for different classes or may depend on the example description. In a loan approval system, where the classes are *applicant will/will not default on loan*, the class *will default on loan* can become known before the end of the loan period, whereas the class *will not default on loan* cannot be determined until the loan is completed.

In addition to latency, examples can be subject to filtering. Trackability filtering refers to the rejection of examples so that they never return to the example base. This is referred to as one-sided feedback [5]. In the fraud domain, transactions classified as fraudulent will be blocked thus no verified class will ever be obtained. An example filtered in this way can also be regarded as having infinite verification latency.

Selection filtering refers to the sampling of examples for verification and, therefore, their subsequent availability to return as training examples. For instance, in an online spam detection system it would be highly unlikely that every email sent would be verified as spam or non-spam.

In general, filtering can be class dependent and may also be dependent on attribute values within the description of an example.

## 2.1 The Online Learning Life Cycle

The online learning life cycle (OLLC) can be modeled to take account of latency and filtering. The key stages are summarised in Table 1 and illustrated in Figure 1.

**Table 1.** OLLC Stages

Stage	Description
Initial example collection	Initial supervised example collection takes place using external sources and is placed into the training base.
Initial classifier induction	Upon receiving data from the training base, the learning algorithm is applied and generates the first classifier.
Classification	A new case, <description>, arrives at time $t_c$ for classification. This is given its predicted class, $pclass$ , by the current classifier and stored as $(t_c, \langle description \rangle, pclass)$ .
Verification and return to the example base	At time, $t_v > t_c$ , the true class, $vclass$ , may be obtained. Verification latency is $t_1^{lat}$ . After further delay, $t_2^{lat}$ , this may be fed back, at $t_{ab}$ , to the example training base as $(\langle description \rangle, vclass)$ .
Learning update regime	The learner is applied periodically to examples returned to the training example base since the last episode of learning.

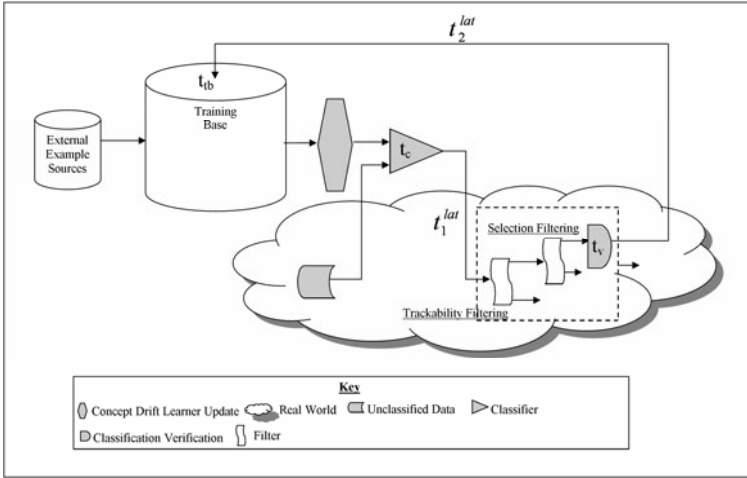


Fig. 1. The OLLC Model

### 3 Experiments on Latency

To investigate the impact of latency on learning under concept drift, a series of experiments was performed using a variety of latency and drift scenarios. In these, latency was modeled as both fixed and random and was assumed to be independent of description attribute values of the examples and of class. No filtering was applied.

#### 3.1 Data Sets

Training and test examples were generated using AutoUniv [6]. AutoUniv creates an artificial universe (U), a complete probabilistic model of the joint distribution of the description attribute and the class, comprising four components: attribute definitions; attribute factorization (into independent factors); attribute factor distributions; and, a rule set. Noise is modelled as the degree of uncertainty in the class distributions of each rule.

Drifted models can be obtained by retaining attribute and class definitions and altering some of the other components of the universe. Drift can be in rules only (concept drift) or in attribute distributions only (population drift) or both.

For the experiments, three universes were generated all with the same attributes and classes, details of which are summarised in table 2. The second universe was obtained by applying concept drift to the first; the third was obtained by applying concept drift to the second. Details are given in tables 3 (a) and 3 (b).

From table 3 (a) all three universes have similar Bayes rate, i.e. the maximum classification accuracy possible. The complement of the Bayes rate is the noise level.



**Table 2.** Universe attributes (common to all drift variations)

Relevant Attributes	Noise Attributes	Attribute Factors	Minimum Number of Attribute Values
8	2	3	2
Maximum Number of Attribute Values	Number of Classes	Minimum Rule Length	Maximum Rules Length
5	4	2	5

**Table 3. (a)** Universe drift variations

Universe	Rules	Average Rule Length	Noise %	Bayes Rate %
1	60	4.2	21.4	78.6
2	88	4.3	19.5	80.5
3	82	4.3	22.1	77.9

**(b)** Cross-classification rates (XCR) between universes

Old	New	XCR (%)
Universe 1	Universe 2	32.1
Universe 2	Universe 3	26.8
Universe 1	Universe 3	33.6

The cross-classification rate (XCR), [7], provides a simple measure of the extent of drift. This is defined as the classification rate that would be obtained if the universe rules operating before a drift point were applied to classify examples after drift had occurred. Intuitively this provides a base-line for the consequences of failing to detect drift. The XCR is bounded above by the Bayes rate for the drifted universe, that is, no rule set can outperform the true rules. From table 3(b) it is seen that XCR is very low, in relation to the corresponding Bayes rate in all cases. If an online learner correctly induces rules for universe 1 but fails to detect the drift to universe 2 and then to universe 3, the classification rate will drop first to about 32% and then rise to about 34%. In practice, the learner will usually not induce a completely correct set of rules for universe 1 and this imperfect rule set could achieve classification rates after drift that are higher than that of the XCR, but still far short of the Bayes rate.

### 3.2 The Learning Algorithm

The CD3 algorithm [7] was selected as the online learner. CD3 uses the decision tree algorithm, ID3, along with post pruning as a base learner. In each episode of learning, CD3 receives in a batch, training examples that have become available since the last episode. These are time-stamped as *new* and added to examples retained from previous episodes, time-stamped as *current*. The time-stamp is added to an example's description prior to induction by ID3. In effect, CD3 is assessing the relevance of the

time-stamp attribute to classification. This is the time-stamp attribute relevance (TSAR) principle [7]. Following induction, rules are extracted from the tree and those which specify the time-stamp value as *current* are deemed to be out of date and purged. Current examples which are covered by a purged rule are purged. Finally, the new examples have their time-stamp changed to *current* for the next learning episode. By this means, CD3 aims to dynamically maintain a base of training examples considered to be valid, that is, they reflect the current oracle. It is an important feature of CD3 that it does remove training examples simply on the basis of age, the argument being that, under concept drift, typically not all rules are subject to drift and so examples covered by un-drifted rules retain their relevance to the learner.

### 3.3 Example Life-Cycle Integration Simulator Environment (ELISE)

Designed to complement examples generated from AutoUniv, ELISE has been developed to load example files into a database and to generate initial time point and additional latency values to include with each example in accordance with the online learner life cycle model in figure 1. Test examples are also loaded into the database.

The system allows for the selection of either constant or random latency types. Random latency offers a further breakdown into latency models generated according to a Normal distribution or negative exponential: a Normal distribution representing latency scenarios where examples may return early or late but will more likely be closer to an expected time, and, negative exponential for the scenario where examples will most likely return soon after classification but still allow for very late example return. When selecting a Normal distribution the user has control over a number of preset variance levels. Overall, random latency is determined from an inputted average latency value.

In addition to the latency periods, the user can specify a regime for examples arriving to be classified. These arrivals can be fixed or random and modeled as above.

ELISE then proceeds to simulate an online environment by handing over batch files of training examples at an inputted time interval, e.g. every 1000 time points ( $t$ ) to a learner, currently CD3. The learner interface, as well as allowing for learning, also has testing and single classification. Upon every learning cycle, the system will call on a learner to test its most recent classifier with a batch of test examples. The learner then writes out its test results to a file.

ELISE requires all initial training examples to be provided as a representation of the first learner supervised example stage. These first examples are given the time point of zero since they represent historic example collection with unknown times. Additional training examples are added through a domain example file. These examples are given chronological initial ID times starting from one and in intervals of one unit. Additional times are then generated and stored to represent the various latency periods the example incurs through the life cycle.

The user inputs the drift points, i.e. the time points at which the first example of each new drifted universe is experienced. They also enter the time point at the end of each test batch and the total number of test examples that make up a batch.

In all, 21 experiments were scripted to run in ELISE. For each of the drift scenarios in table 4, each latency scenario in table 5 was conducted for both a medium and high latency value, as presented in table 6. The first experiment in each drift scenario, i.e. zero latency experiment, was performed to determine CD3 baseline performance under each drift scenario prior to the addition of latency. This made possible an assessment of the nature of and extent to which latency impacted upon the various drift scenarios. Ten iterations of each of these experiments were performed using different sets of data taken from each drifted universe.

**Table 4.** Drift scenarios

Drifts	Drift points
0	0
1	4501
2	3001, 6001

**Table 5.** Latency scenarios

Latency Type	Latency Model
Zero	n/a
Constant	n/a
Random	Normal Distribution
Random	Negative Exponential

**Table 6.** Standard deviations for random latency

Average Latency	Normal Distribution Standard Deviation	Negative Exponential Standard Deviation
500	96	500
2000	516	2000

Data was supplied as described in table 7. It should be noted that the combined total of examples, 10000, used for learning never changes in the experiments. The effect of increasing the number of drifts reduces the overall time for recovery and therefore represents an increasingly drift-active domain. Also, these initial experiments only consider the impact of latency on domains susceptible to revolutionary drift, i.e. a sudden and immediate change in the rules. This gives a clearer interpretation of the impact of latency. However, it is intended that evolutionary drift, i.e. gradual change in rules, will also be explored in later latency experiments.

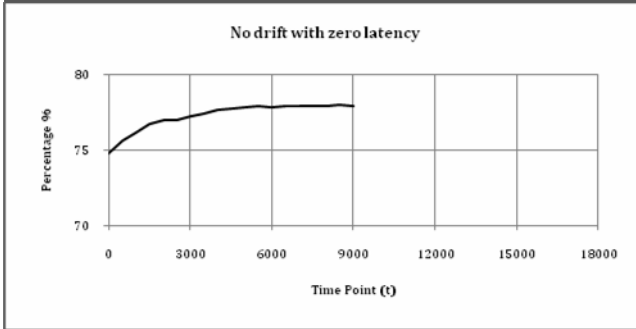
**Table 7.** Example data artificial universe(U) breakdown for each drift scenario

Category	Training Examples	Domain Examples	Test Examples
No drift	U1 (1000)	U1(9000)	U1(10000)
One drift	U1 (1000)	U1(4500), U2(4500)	U1(10000), Universe 2(10000)
Two drift	U1 (1000)	U1(3000), U2(3000), U3(3000)	U1(10000), U2(10000), U3(10000)

Finally, learning and testing cycles were performed in the ELISE simulation every 500 time points.

## 4 Analysis of Results

The initial experiments involving no drift performed as expected. With the learner commencing at time point zero, having already conducted an initial learning run of 1000 training examples, it quickly achieved a high classification accuracy bordering on the Bayes rate for the universe, as in figure 2. Under latency, the only difference was in the return of examples after the last true time point of 9001, indicating that the examples were returning later than their true time, i.e. lagging as the result of latency.



**Fig. 2.** Test result for no drift with zero latency experiment

The subsequent one drift experiments demonstrate interesting latency rate impacts upon the learner's ability to provide accurate classifiers in time for drift. With zero latency (see figures 3 and 4), the classification rate (CR) crashes at the drift point and then immediately recovers. As seen in comparison to table 8, the lowest accuracy in each of the latency and drift scenarios is comparable to the cross-classification rate (XCR). However, the pattern of recovery is entirely different under latency conditions.

Under the medium latency experiments shown in figure 3, it can be seen that for each of the three latencies a delay is incurred prior to recovery being made. This impact is even more pronounced under high latency.

For constant latency, a low classification plateau occurs for the duration of the example latency, i.e. the nature of constant latency presents as being an overall constant lag behind the current domain. In the high latency scenario, the example latency was for 2000 time points and this matches the duration of the lag prior to recovery. This highlights issues in selecting the time stamp to represent an example in online learners.

While recovery under constant latency appears to be quicker, the Normal distribution performs similarly, although slightly behind. It would appear that the benefit of earlier examples from the new universe is counteracted by the possibility of old examples from the previous. However, a near Bayes rate classification accuracy is achieved by the end of the last true domain time point of 9001.

Negative exponential latency provides the most interesting results. While initially beginning its recovery quicker than under constant latency, the overall rate of recovery is much slower than the other latency models with a failure to achieve anywhere

near the Bayes rate. The overall impact of these patterns becomes even more pronounced and obvious when examined in the two drift experiment results as in figure 4.

In this instance it is seen that even without latency the learner is beginning to struggle by the second drift: failing to achieve a near Bayes rate classification by the last example's true domain time point of 9001. The reduction in time between each drift sees a compounding realisation in the classification rate crash and recovery with the second drift point occurring prior to full recovery.

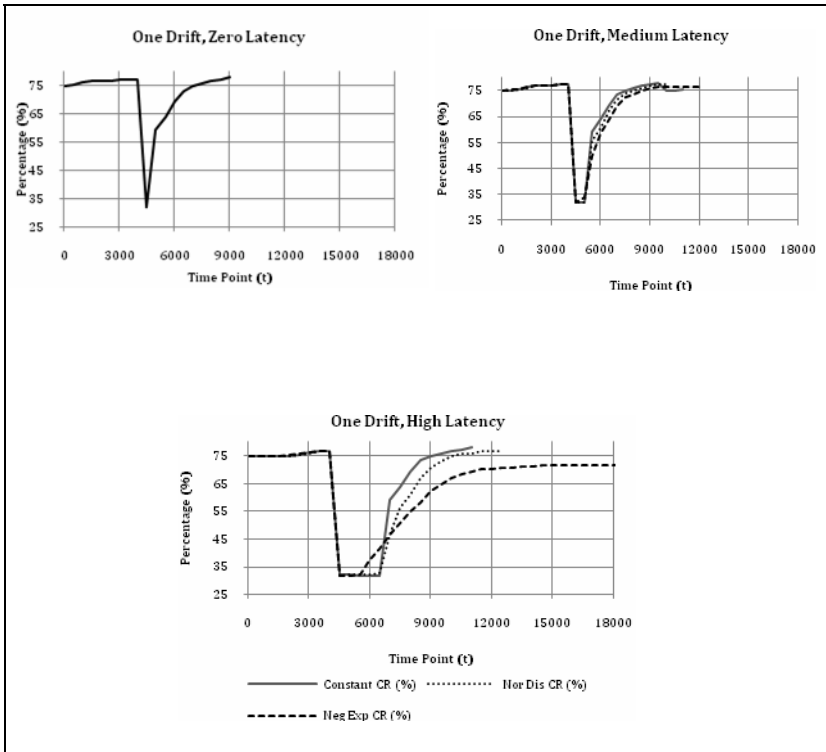


Fig. 3. Test results for one drift scenario experiments

The latency models provide further revelations as to their impact on an online learner under the two drift scenario. While it can be seen that the effect observed under the one drift experiments for both constant and Normal distribution latency is exaggerated further, the extent of damage that a negative exponential latency has upon a learner's classification accuracy becomes more apparent.

The first drift under negative exponential fails to recover in any useful way before the second drift occurs. This results in the second rate of recovery attempt being even lower and stabilising at only 54% classification accuracy around time point 17000 therefore never recovering: 23.9% less than the Bayes rate. In fact, it is only after time point 22001 that the final examples return.

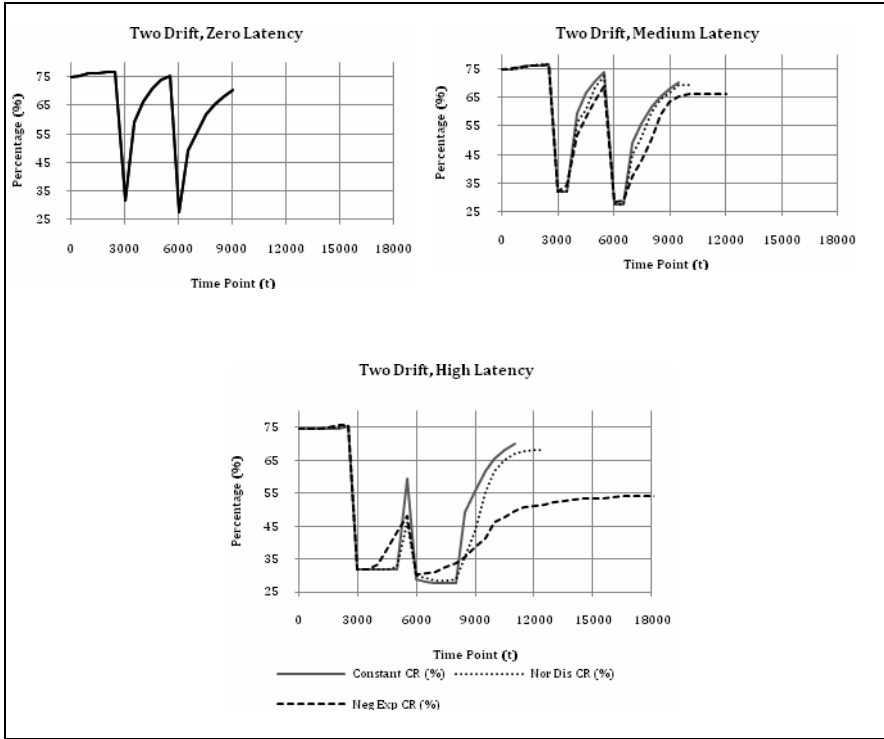


Fig. 4. Test results for two drift scenario experiments

Upon further consideration, the cause of the negative exponential latency’s severe impact upon the learner is clear. At the first drift point the learner is still receiving late examples from universe one delaying its recovery. By the second drift point, the learner is still receiving examples from both universe 1 and universe 2 in addition to the new universe. See table 8 where the composition of the first two batches after the drift point are displayed. The contamination caused from the mixing of the universes presents itself to the learner as a new compounded pseudo-universe.

Negative exponential latency in this experiment gives a mixture of examples that are not only contaminated but that also have a majority representation from a previous universe. As a result, it is not possible for the learner to achieve a successful classifier.

Table 8. Example batch composition for two drift, high latency experiment

Batch Times (t)	Average Percentage %		
	Universe 1	Universe 2	Universe 3
6001 - 6501	16	72	12
6501 - 7001	12.7	55.7	31.6

## 5 Conclusion and Future Work

Latency in training examples, as defined here, has been shown to have a marked impact on the ability of the online learner CD3 to recover from concept drift. It is reasonable to suppose that other online learners are likely to be similarly affected. Systems in which there is capacity for large latencies such as modelled here by the negative exponential distribution are especially vulnerable. It can therefore be argued that any online learner subject to latency should be tested for accuracy and recovery under these various latency models prior to being deployed.

Further experiments are planned to investigate the performance of an online learner under latency involving different drift scenarios including evolutionary drift as well as under class / attribute - specific example filtering.

Meta attributes such as classification and verification times and latency itself were defined here in an example life cycle for online learning. The resulting meta data can be retained in the training base. The statistical information it provides will enable profiling of the data stream and the training base in addition to directly supporting learning.

As a first step towards equipping an online learner to handle latency, CD3 will be augmented to make use of latency as a meta or context attribute in learning.

## References

1. Kolter, J.Z., Maloof, M.A.: Dynamic Weighted Majority: An ensemble method for drifting concepts. *Journal of Machine Learning Research* 8, 2755–2790 (2007)
2. Minku, L.L., White, A.P., Yao, X.: The Impact of Diversity on On-line Ensemble Learning in the Presence of Concept Drift. *IEEE Transactions on Knowledge and Data Engineering*, 730–742 (2009)
3. Gao, J., Fan, W., Han, J.: On appropriate assumptions to mine data streams: Analysis and practice. In: Perner, P. (ed.) *ICDM 2007*. LNCS (LNAI), vol. 4597, pp. 143–152. Springer, Heidelberg (2007)
4. Wang, H., Yin, J., Pei, J., Yu, P., Yu, J.: Suppressing model over-fitting in mining concept-drifting data streams. In: *Proc. KDD 2006*, Philadelphia, August 20–23, pp. 736–741 (2006)
5. Sculley, D.: Practical learning from one-sided feed-back. In: *Proc. of the 13th ACM SIGKDD international conference on Knowledge discovery and data mining, KDD 2007* (2007)
6. Hickey, R.J.: Structure and Majority Classes in Decision Tree Learning. *Journal of Machine Learning* 8, 1747–1768 (2007)
7. Black, M., Hickey, R.J.: Maintaining the performance of a learned classifier under concept drift. *Intelligent Data Analysis* 3, 453–474 (1999)

# Efficient Reasoning with RCC-3D

Julia Albath, Jennifer L. Leopold, Chaman L. Sabharwal, and Kenneth Perry

Missouri University of Science and Technology  
Department of Computer Science  
Rolla, MO 6540, USA  
{jgadc, leopoldj, chaman, klpwdf}@mst.edu

**Abstract.** Qualitative spatial reasoning is an important function of the human brain. Artificial systems that can perform such reasoning have many applications such as Geographic Information Systems (GIS), robotics, biomedicine, and engineering. Automation of such analytical processes alleviates manual labor, and may increase the accuracy of the spatial assessments because the reasoning can be done objectively using 3D digital representations of the objects. Herein we introduce an algorithm to determine the spatial relation that exists between a pair of 3D objects when no a priori spatial knowledge is given. A second algorithm is presented to efficiently find the spatial relation that holds between each pair of objects in a set of 3D objects.

**Keywords:** Spatial reasoning, knowledge representation, knowledge-based system, relational connection calculus.

## 1 Introduction

Spatial reasoning is an important function of the human brain. When driving on a highway we do not need to know that another car is approaching our vehicle at 63.4 mph, but rather it is enough to know that it is approaching too rapidly to merge safely. In such situations we do not require quantitative information; qualitative information is sufficient. Qualitative Spatial Reasoning (QSR) research addresses the issues related to automating such analytical processes [1].

There are many applications of QSR, from Geographic Information Systems (GIS), to robot navigation, biology, medicine, and engineering problems [2]. For example, a morphologist may need to examine and describe a species with respect to the dorsoventral (front to back), anteroposterior (head to toe), and left-right axes. These morphological descriptions adhere to informal standards, yet they must painstakingly be performed manually by a trained biologist. Given the escalating rate of new species discovery and the backlog of undescribed specimens in natural history museum collections, this time-consuming process continues to create a huge bottleneck in critical biodiversity research.

Today many types of 3D data can be generated easily, inexpensively, and unobtrusively, including MRI and CT scans of organism specimens; however, such



datasets can be fairly large in terms of the number of data points. RCC-3D was introduced as a model that facilitates spatial reasoning over 3D digital data with consideration of perspective [3]. Herein we introduce algorithms that can efficiently determine the RCC-3D relation that holds between two 3D objects, and between each pair of objects in a collection of 3D objects.

## 2 Related Works

The actual incorporation of a spatial calculus into a domain specific application can be challenging; the performance of implemented calculi can vary, and it is difficult to know prior to implementation which calculi will perform best for a particular problem domain. The SparQ toolkit was created to facilitate development of QSR applications [4]. It includes a qualifier module for various calculi to compute the spatial relation between two or more objects. However, such modules have not been created for all region connection calculi; for example, in order to use SparQ with RCC-8 [5], one must provide input in the form of “(A EQ B) (B TPP C)”, where A, B, and C are distinct objects, and EQ and TPP represent the RCC-8 relations for equality and tangential proper overlap, respectively. In contrast, the QSR algorithms proposed in this paper actually determine the spatial relation that holds between each pair of objects; that is, our algorithms have implemented the abstraction from a quantitative to a qualitative model.

Another use of QSR is the study of qualitative simulations using a constraint-based framework [6]. Linear temporal logic is used to create a constraint programming system (CPS) that deals with time, space, shape, and size. Unlike our algorithms, the focus is not in determining if the relations among the objects are consistent, but rather if the CPS itself is consistent. For example, a CPS could be used to simulate a person juggling three balls; a sequence of relations would specify the spatial constraints that must hold at each temporal stage to create an infinite juggling simulation. Input to this application would include information such as “Q[left-hand, right-hand] = disjoint” for a particular moment of time that is to be modeled. In contrast, our algorithms effectively use abstractions of the real world (images) to actually determine the set of spatial relations that hold at a given time.

Geographic Information Systems often utilize QSR. For example, [7] describes the use of a QSR system to describe inconsistencies, such as a newly constructed road that is not present in an older geographic dataset. Such applications often use description logic, with the goal of integrating different datasets. Again, that use differs from our more general objectives, which are to efficiently determine the spatial relations that hold between objects while ensuring consistency.

## 3 Terminology and an Overview of Relations

QSR theories are based on a Jointly Exhaustive and Pairwise Disjoint (JEPD) set of relations [8]. JEPD implies that each pair of objects can and must belong to exactly

one relation. RCC-3D is based on the Generalized Region Connection Calculus (GRCC) [9], which provides the same eight relations as RCC-8 [5]. However, RCC-3D includes additional relations to account for the parallel projections that are perpendicular to each of the principal planes in  $R^3$  (i.e., the planes through the xy-axes, yz-axes, and zx-axes).

The fundamental axioms of RCC-3D are *Parthood* and *Connectivity*. We use  $A \cap B$  to represent everything common to both A and B.  $A \sim B$  is used to denote everything that is in A, but not in B.  $A^e$  is the exterior of A; it is used to represent everything that is not in the closure of A (which is denoted as  $\overline{A}$ ). The boundary of A is denoted by  $\partial A$ ; hence, the closure of A is everything in A and  $\partial A$ .  $A^\circ$  denotes the interior of A, which is everything that is in the closure of A, but does not include the boundary of A. The interior, boundary, and exterior of an object are disjoint, and their union is the universe. We define  $P(A,B)$  by saying that A is part of B if A is a subset of B; specifically, it must be the case that  $A \cap \overline{B^c} = \emptyset$ , where  $B^c$  is the complement of B. We define  $C(A, B)$  as A is connected to B if  $\overline{A} \cap \overline{B} \neq \emptyset$ .

A predicate, *Proper Part (PP)*, is derived from  $P$ :

$$PP(A,B) \equiv P(A,B) \wedge \neg P(B,A) \quad (1)$$

There are eight JEPD relations that are used to form the foundation of the RCC-3D theory, all of which can be defined in terms of  $P$ ,  $C$ , and  $PP$ :

$$PO(A,B) \equiv \exists (C \neq \emptyset) [P(C,A^\circ) \wedge P(C,B^\circ)] \wedge \neg P(A,B) \wedge \neg P(B,A) \quad (2)$$

$$EQ(A,B) \equiv A=B \quad (3)$$

$$EC(A,B) \equiv C(A,B) \wedge (A^\circ \cap B^\circ = \emptyset) \quad (4)$$

$$DC(A,B) \equiv \neg C(A,B) \quad (5)$$

$$TPP(A,B) \equiv PP(A,B) \wedge \exists (C \neq \emptyset) [EC(C,A^\circ) \wedge EC(C,B^\circ)] \quad (6)$$

$$NTPP(A,B) \equiv PP(A,B) \wedge \exists C [EC(C,A) \wedge EC(C,B)] \quad (7)$$

$$TPP_c(A,B) \equiv PP(B,A) \wedge \exists (C \neq \emptyset) [EC(C,B^\circ) \wedge EC(C,A^\circ)] \quad (8)$$

$$NTPP_c(A,B) \equiv PP(B,A) \wedge \exists C [EC(C,B) \wedge EC(C,A)] \quad (9)$$

If an object X is in relation  $R$  to object Y, then the object Y is in *converse relation*, denoted by  $R_c$ , to object X. For example,  $TPP_c$  is the converse of  $TPP$ . For symmetric relations, there is no need to suffix  $c$  to the relation name (i.e., the relations  $DC(A,B)$ ,  $EC(A,B)$ ,  $EQ(A,B)$ , and  $PO(A,B)$  are symmetric, and do not require distinctly named converse relations).

For some types of spatial reasoning, such as anatomy, it is necessary to consider the obscuration that can occur when the objects are seen through orthogonal projection on any of the principal planes in  $R^3$ . When considering that object A obscures object B, it is implied that object A is closer than object B to the perspective reference point. For our discussion, we assume that the direction of the projection (i.e., the line of sight) is orthogonal to the plane of projection, and that the plane of projection is one of the principal planes.

Let  $A_p$  and  $B_p$  be the projections of A and B on a plane P, where P is  $xy$ ,  $yz$ , or  $zx$ . The knowledge of the order of objects is discernible from the 3D data. For  $EQ$ ,  $TPP$ , and  $NTPP$  obscuration is implied. The relations that do require qualification in terms of obscuration (and, hence, distinguish RCC-3D from RCC-8) are listed below.

$$DC_{pp}(A,B) \equiv DC(A,B) \wedge PO(A_p,B_p) \quad (10)$$

$$DC_p(A,B) \equiv DC(A,B) \wedge (TPP_c(A_p,B_p) \vee NTPP_c(A_p,B_p)) \quad (11)$$

$$EC_{pp}(A,B) \equiv EC(A,B) \wedge PO(A_p,B_p) \quad (12)$$

$$EC_p(A,B) \equiv EC(A,B) \wedge (TPP_c(A_p,B_p) \vee NTPP_c(A_p,B_p)) \quad (13)$$

$$PO_{pp}(A,B) \equiv PO(A,B) \wedge PO(A_p,B_p) \quad (14)$$

$$PO_p(A,B) \equiv PO(A,B) \wedge (TPP_c(A_p,B_p) \vee NTPP_c(A_p,B_p)) \quad (15)$$

Each of the relations that consider obscuration (Equations (10)-(15)) has a converse, which is distinguished by adding  $c$  to the end of the relation name.

Additionally, RCC-3D differentiates between *complete* and *partial* obscuration. Partial obscuration is indicated by the inclusion of  $p$  in the relation name; for example,  $DC_{pp}(A,B)$  is true if objects A and B are disconnected and A partially obscures B in the P plane, whereas  $DC_p(A,B)$  is true if objects A and B are disconnected and A completely obscures B in the P plane. This characterization is critical for some problem domains such as medical diagnostics and engineering.

Vast amounts of digital 3D datasets are available today. Yet they rarely are used to their full potential to support humans in various reasoning tasks. For large collections of 3D objects, it may not be feasible to calculate (and permanently record) all the spatial relationships; furthermore those relationships between objects may be subject to repeated modification, as in a mechanical design. Thus, it is important to have an efficient algorithm that can make such determinations in both a sufficiently descriptive and logically consistent manner.

## 4 Algorithms

RCC-3D has been implemented as a C++ program to facilitate qualitative spatial reasoning without any *a priori* knowledge about the underlying relations. The data used by RCC-3D are digital 3D objects, which have been marked mechanically with boundary information. An object is defined in terms of faces and face-defining vertices. The RCC-3D reasoning system uses these data to quantitatively determine the

truth-values of the necessary predicates; namely, geometric calculations are performed over the faces and face-defining vertices for each pair of objects. Those predicates then are used to determine which RCC-3D (qualitative) relation applies to the objects under consideration. An existing constraint solver tool (e.g., MINION [10]) was not utilized for our prototype implementation of the system due to the anticipated computational demands of the geometric calculations; however, such an approach may be investigated in the future.

The cost to calculate each predicate varies considerably. To determine a particular RCC-3D relation, only a certain combination of predicates is required. Some relations have multiple combinations that can be used. Hence performing the calculations in an order that will facilitate efficient calculation of a large number of RCC-3D relations is desirable.

#### 4.1 Computational Cost, Efficiency, and Predicates

The 9-Intersection Calculus [11] classifies relations by the truth values of the intersection of A's interior, boundary, and exterior with B's interior, boundary, and exterior. The 512 possible intersections are reduced through negative relations representing impossible situations (i.e., it would be impossible to have A's interior intersect B's interior and B's exterior, but not B's boundary). We apply a similar strategy by defining five predicates (and a converse) to uniquely identify the RCC-3D relations. As each predicate is calculated, we can reduce the set of possible relations that can hold between the two objects of interest by avoiding inconsistencies.

The computational cost of calculating each predicate is related to the size of the two objects A and B; namely, it depends on the number of faces ( $f_A$  and  $f_B$ ) and the number of vertices ( $v_A$  and  $v_B$ ). The value of each of these parameters is known after reading the 3D image data file for an object. When there is more than one combination of predicates that can be used to identify an RCC-3D relation, we use the combination of predicates that has the lowest total cost. It should be noted that the complexity given for each predicate below is with respect to our current implementation of RCC-3D; however, the definition of each predicate is independent of the implementation.

In this discussion it is assumed that each predicate takes two arguments, objects A and B, and returns a Boolean result. The first predicate is Boundary-Boundary-Test  $BBT(A,B)(\partial A \cap \partial B \neq \emptyset)$ , which determines whether the boundaries of the two objects intersect. It is used to distinguish between *TPP* and *NTPP*, between *TPPc* and *NTPPc*, and between *DC* and *EC*. Because there are at most a fixed number of vertices at each face and there are at most a fixed number of faces at each vertex, we assume that  $O(v_A) = O(f_A)$ . To simplify the representation (and comparison) of the complexity of the predicates, let  $n = \max(|f_A|, |f_B|, |v_A|, |v_B|)$ . The cost of calculating  $BBT(A,B)$  is then  $O(n^2)$ .

Next we consider the predicate Exterior-Interior-Test  $EIT(A,B) (A^e \cap B^o \neq \emptyset)$  and its converse  $EITc(A,B): A^o \cap B^e \equiv B^e \cap A^o (A^o \cap B^e \neq \emptyset)$ . These are used to determine if the exterior of one object intersects with the interior of the other object;

that is, EIT tests whether any part of A’s exterior is outside the boundary of B. If EIT is FALSE, then  $B^\circ$  must be contained within A. When EIT is TRUE, we know that B is not a proper subset of A. The cost of calculating both  $EIT(A,B)$  and  $EITc(A,B)$  is  $O(n^3)$ .

The next predicate is Interior-Interior-Test  $IIT(A,B) (A^\circ \cap B^\circ \neq \emptyset)$ , which is used to determine whether the interiors of two objects overlap. If we only used the BBT, EIT, and EITc tests, there would be no way to distinguish between *EC* and *PO*; when IIT also is considered, those distinctions are possible. The cost of calculating  $IIT(A,B)$  is  $O(n^5)$ .

The last two predicates are used to differentiate between full and partial obscuration in one of the principal planes. To determine obscuration in plane P, we use  $EIT_p(A,B) (A_p^e \cap B_p^o \neq \emptyset)$  and  $IIT_p(A,B) (A_p^\circ \cap B_p^\circ \neq \emptyset)$ . If  $A_p$ ’s exterior does not intersect  $B_p$ ’s interior (EIT is FALSE), then A *fully* obscures B. If  $A_p$ ’s interior intersects  $B_p$ ’s interior and  $A_p$ ’s exterior intersects  $B_p$ ’s interior, then A *partially* obscures B. The costs for  $EIT_p(A,B)$  and  $IIT_p(A,B)$  are both  $O(n^3)$ .

Table 1 characterizes each RCC-3D relation with regards to the aforementioned predicates. An entry in column  $k$ , where  $k$  is one of the predicates, is T if the answer to the question “is it true that  $k(A,B)$ ” is yes. If the answer to that question is no, the table entry in that column is F. A table entry denoted with ‘-’ represents a “don’t care” value; that is, the RCC-3D relation can be determined without considering that particular predicate value. The table entries are ordered from left to right, and from top to bottom, in terms of increasing computational complexity of the predicates, and with T entries before F entries.

**Table 1.** The 13 RCC-3D relations characterized by the five predicates and one converse predicate

BBT	EIT	EITc	$EIT_p$	$IIT_p$	IIT	Relation
T	T	T	T	T	T	$PO_{pp}$
T	T	T	T	T	F	$EC_{pp}$
T	T	T	T	F	-	EC
T	T	T	F	-	T	$PO_p$
T	T	T	F	-	F	$EC_p$
T	T	F	-	-	-	$TPP_p$
T	F	T	-	-	-	$TPP_{pC}$
T	F	F	-	-	-	$EQ_p$
F	T	T	T	T	-	$DC_{pp}$
F	T	T	T	F	-	DC
F	T	T	F	-	-	$DC_p$
F	T	F	-	-	-	$NTPP_p$
F	F	-	-	-	-	$NTPP_{pC}$

For a large collection of objects, it should not be necessary (and may not be efficient) to submit a separate query for the spatial relation that holds between a pair of objects, for each pair in the collection. Henceforth we provide two algorithms. The



It should be noted that the structure of the decision tree utilized for Algorithm 1 is based on the computational complexity of the predicates (as specified in Section 4.1) instead of the information gain that would be achieved at each branch.

### 4.3 Multiple Objects

There are applications for which a collection of more than two 3D objects is available, and the RCC-3D relation that holds between each pair of objects in that collection is required. Systematically executing the Single-Pair-Relation-Detection Algorithm for each pair of objects in the collection is not the most efficient manner in which to proceed. Instead we introduce an All-Pairs-Relation-Detection Algorithm, which utilizes Algorithm 1 in conjunction with a Path-Consistency (PC) Algorithm.

Let  $N$  be the number of 3D objects in the collection. Algorithm 2 uses an  $N \times N$  table  $R$  where  $R[X, Y]$ ,  $1 \leq X, Y \leq N$ , contains a data structure with fields for: (1) the set of RCC-3D relations that possibly could hold between objects  $X$  and  $Y$ , and (2) the total cost to compute all of the six predicates listed in Table 1 (i.e., BBT, EIT, EITc, EITp, IITp, and IIT) for objects  $X$  and  $Y$ . After initializing the entries in this table, the algorithm loops until the RCC-3D relation that holds between each pair of objects in the collection has been uniquely determined. Each iteration of the loop chooses a pair of objects  $X$  and  $Y$  for which the (single) RCC-3D relation between the objects has yet to be determined, and for which the total cost of computing all six predicates is the lowest. Algorithm 1 then is used to determine the RCC-3D relation that holds between objects  $X$  and  $Y$  (as well as the relation that holds between  $Y$  and  $X$ ). This information subsequently is used by Algorithm 3 to determine what else can be deduced about the set of possible relations that can hold for the remaining pairs of objects.

The PC Algorithm, shown in Algorithm 3, is modeled after the dynamic PC algorithm presented in [12]. This algorithm takes as input the object pair for which Algorithm 1 has just calculated the single RCC-3D relation that holds, and uses the composition table for the base relations to calculate the set of the possible relations for  $(A, C)$ , given what is known about  $(A, B)$  and  $(B, C)$ . When this set is different from what is currently known about  $(A, C)$ , we update  $R[A, C]$ . Whenever the PC algorithm updates the set of possible relations for an object pair, that pair is appended to an 'updated' list. Before an object pair is processed in this manner it is removed from the 'updated' list. The algorithm continues until the 'updated' list is empty.

Each iteration of the All-Pairs-Relation-Detection Algorithm uses the Path-Consistency Algorithm to update the relation information for other pairs of objects (the RCC-3D relation for which has yet to be determined). While we intend to create the composition table for the RCC-3D calculus as part of a future work, we are using the composition table presented in [13] for the RCC-8 relations (which are the basis for the RCC-3D relations) until such time; consideration of obscuration is done after the composition table is referenced to determine the base relation. The Path-Consistency Algorithm reduces the search space of possible relations, and improves the efficiency of the All-Pairs-Relation-Detection Algorithm.





```

while L is not empty do
  (X,Y) = head of L;
  remove (X,Y) from L

  // Find the set M of possible relations for the object pair (X,K) using the table CT
  M = { };
  for K = 1 to N, K != X and K != Y do
    S = R[X,Y].possibleBaseRelations;
    Q = R[Y,K].possibleBaseRelations;
    for i = 1 to S.size, j = 1 to Q.size do
      M = M  $\cup$  CT[Si,Qj]
    end-for
  T = R[X,K].possibleRelations  $\cap$  M;
  // If T is different from R[X,K].possibleRelations, update R[X,K].possibleRelations with
  // T, and update R[K,X].possibleRelations with the converse of each relation in T
  if (T != R[X,K].possibleRelations)
    begin
      R[X,K].possibleRelations = T;
      R[K,X].possibleRelations = Tc , the converse of each relation in T;
      L = L  $\cup$  {(X,K)}  $\cup$  {(K,X)};
      // Calculate obscuration if T is a singleton set, and that one relation is PO, DC, or EC
      if (T.size() == 1 && (T == PO || EC || DC))
        begin
          if (EITp(X,K) == TRUE && IITp(X,K) == TRUE)
            R[X,K].possibleRelations = Tpp
          else begin
            if (EITp(X,K) == FALSE)
              R[X,K].possibleRelations = Tp
            end
          if (EITp(K,X) == TRUE && IITp(K,X) == TRUE)
            R[K,X].possibleRelations = Tcp
          else begin
            if (EITp(K,X) == FALSE) R[K,X].possibleRelations = Tcp
          end
        end
      end
    end-for
  end-while

```

As a very simplified example of the All-Pairs-Relation-Detection Algorithm, suppose we have three objects, A, B, and C, and we only are considering the eight base relations (Equations (2)-(9)). Assume that, based on the predicate calculation costs, we determine that the pair (A,B) is to be considered first in Algorithm 2. After *NTPP* is calculated for (A,B), Algorithm 3 subsequently determines that relation information for the pairs (A,C) and (C,A) can be refined; namely, that  $R[A,C].possibleRelations$  and  $R[C,A].possibleRelations$  can both become  $\{EC, DC, PO, TPP, NTPP, NTPPc, TPPc\}$ . Next Algorithm 2 uses Algorithm 1 to determine the (single) relation for the pair (B,C), and, by extension, the (single) relation for the pair (C,B). Algorithm 3 then updates possibleRelations for  $R[A,C]$  and  $R[C,A]$ . After Algorithm 3 has removed (C,B) from list L, it is determined that object K must be A; consequently,  $S = (C,B)$ ,  $Q = (B,A)$ , and  $M = CT[TPPc, NTPPc] = \{NTPPc\}$ . After performing the intersection

with possibleRelations for  $R[C,A]$ , possibleRelations for  $R[C,A]$  and  $R[A,C]$  are updated to  $\{NTPP_c\}$  and  $\{NTPP\}$ , respectively. Table 2 shows the completed table R after these updates.

**Table 2.** Final relations-table for the example

Object	A	B	C
A	$\{EQ\}$	$\{NTPP\}$	$\{NTPP\}$
B	$\{NTPP_c\}$	$\{EQ\}$	$\{TPP\}$
C	$\{NTPP_c\}$	$\{TPP_c\}$	$\{EQ\}$

## 5 Experimental Results

To determine the relationships between 52 3D objects that represent a human brain (images obtained from [14]), our current implementation of the All-Pairs-Relation-Detection Algorithm took approximately eight hours on an Intel(R) Core(TM)2 Quad CPU Q6600 @ 2.40GHz with 8GB of RAM computer running Kubuntu Linux 9.10 (Linux 2.6.31). This dataset had a total of 16014 faces and 314330 vertices. The average object had 3079 faces and 6045 vertices. The largest object had 82104 faces and 160684 vertices; the smallest object had 120 faces and 238 vertices. The algorithm gave accurate and complete results for our application domain.

A parallelized implementation of the algorithm was executed for the same dataset. Only two hours were required to make the same RCC-3D relation determinations on a cluster of computers, using 32 Intel(R) Xeon(TM) CPU 3.20GHz running RedHat Enterprise Linux 5, each machine in the cluster having 2G of RAM. The significant reduction in execution time using parallelized versions of the algorithms clearly increases the feasibility of using the algorithms for large collections of 3D objects.

## 6 Conclusion

Qualitative spatial reasoning has many applications in Geographic Information Systems, biomedicine, and engineering. To fully realize the potential of 3D qualitative spatial reasoning systems, algorithms for determining the spatial relation that holds between two objects must be available. And, because 3D objects may be defined in terms of thousands of vertices, those algorithms must be as efficient as possible.

Herein we introduced an algorithm to efficiently determine the spatial relation that exists between a pair of 3D objects when no *a priori* spatial knowledge is given. A second algorithm was presented to perform this function for a collection of 3D objects, finding the spatial relation that holds between each pair of objects. The latter algorithm utilizes a path-completion algorithm for efficient detection and reasoning.

For future work, we plan to test our software on datasets from different problem domains, such as 3D mechanical diagrams and architectural data, to investigate whether other considerations should be incorporated into the algorithms; for example, should the transparency or translucency of an object A be considered when testing whether it obscures another object B. We believe that continued testing and refinement of this work could significantly enhance both spatial knowledge representation and reasoning.

## Acknowledgements

This work was supported by NSF under awards DBI-0445752 and DBI-0640053.

## References

1. Escrig, M.T., Toledo, F.: *Qualitative spatial Reasoning: Theory and Practice: Application to Robot Navigation*. IOS Press, Amsterdam (1998)
2. Cohn, A.G., Renz, J.: *Qualitative Spatial Representation and Reasoning*. In: *Handbook of Knowledge Representation*, pp. 551–596. Elsevier, Amsterdam (2008)
3. Albath, J., Leopold, J.L., Sabharwal, C.L., Maglia, A.M.: *RCC-3D: Qualitative Spatial Reasoning in 3D*. Submitted to *Journal of Biomedical Semantics* (2009)
4. Randell, D.A., Cui, Z., Cohn, A., Nebel, B., Rich, C., Swartout, W.: *A spatial logic based on regions and connection*. In: *Proceedings of the Third International Conference on Principles of Knowledge Representation and Reasoning, KR 1992*, pp. 165–176. Morgan Kaufmann, San Francisco (1992)
5. Wallgrün, J., Frommberger, L., Wolter, D., Dylla, F., Freksa, C.: *Qualitative spatial representation and reasoning in the sparq-toolbox*. In: Barkowsky, T., Knauff, M., Ligozat, G., Montello, D.R. (eds.) *Spatial Cognition 2007*. LNCS (LNAI), vol. 4387, pp. 39–58. Springer, Heidelberg (2007)
6. Apt, K.R., Brand, S.: *Infinite Qualitative Simulations by Means of Constraint Programming*. In: Benhamou, F. (ed.) *CP 2006*. LNCS, vol. 4204, pp. 29–43. Springer, Heidelberg (2006)
7. Duckham, M., Lingham, J., Mason, K., Worboys, M.: *Qualitative Reasoning about Consistency in Geographic Information*. *Information Sciences* 176(6), 601–627 (2006)
8. Ligozat, G., Renz, J.: *What is a qualitative calculus? A general framework*. In: Zhang, C., Guesgen, H.W., Yeap, W.-K. (eds.) *PRICAI 2004*. LNCS (LNAI), vol. 3157, pp. 53–64. Springer, Heidelberg (2004)
9. Li, S., Ying, M.: *Generalized Region Connection Calculus*. *Artificial Intelligence* 160(1-2), 1–34 (2004)
10. Gent, I., Miguel, I., Rendl, A.: *Tailoring Solver-Independent Constraint Models: A Case Study with Essence' and Minion*. In: Miguel, I., Ruml, W. (eds.) *SARA 2007*. LNCS (LNAI), vol. 4612, pp. 184–199. Springer, Heidelberg (2007)
11. Egenhofer, M.J., Herring, J.: *Categorizing binary topological relations between regions, lines, and points in geographic databases*. *NCGIA Technical Reports*, 91-7 (1991)
12. Mouhoub, M.: *Dynamic path consistency for interval-based temporal reasoning*. In: *Proceedings of the IASTED 21st International Conference on Artificial Intelligence and Applications*. ACTA Press (2003)
13. Renz, J., Nebel, B.: *Spatial Reasoning with Topological Information*. In: Freksa, C., Habel, C., Wender, K.F. (eds.) *Spatial Cognition 1998*. LNCS (LNAI), vol. 1404, pp. 351–372. Springer, Heidelberg (1998)
14. Zygote Media Group: <http://www.zygote.com>

# Automated Ontology Generation Using Spatial Reasoning

Alton Coalter and Jennifer L. Leopold

Missouri University of Science and Technology  
Department of Computer Science  
1870 Miner Circle, Rolla, Missouri 65409, USA  
{abcp7c, leopoldj}@mst.edu

**Abstract.** Recently there has been much interest in using ontologies to facilitate knowledge representation, integration, and reasoning. Correspondingly, the extent of the information embodied by an ontology is increasing beyond the conventional *is\_a* and *part\_of* relationships. To address these requirements, a vast amount of digitally available information may need to be considered when building ontologies, prompting a desire for software tools to automate at least part of the process. The main efforts in this direction have involved textual information retrieval and extraction methods. For some domains extension of the basic relationships could be enhanced further by the analysis of 2D and/or 3D images. For this type of media, image processing algorithms are more appropriate than textual analysis methods. Herein we present an algorithm that, given a collection of 3D image files, utilizes Qualitative Spatial Reasoning (QSR) to automate the creation of an ontology for the objects represented by the images, relating the objects in terms of *is\_a* and *part\_of* relationships and also through unambiguous Relational Connection Calculus (RCC) relations.

**Keywords:** Automated ontology construction, relational connection calculus, qualitative spatial reasoning, knowledge representation, knowledge-based system.

## 1 Introduction

As the utilization of ontologies becomes more prevalent in today's scientific and computing communities, there has been an increase not only in the number of ontologies that exist, but also in the size and complexity of the ontologies; that is, many ontologies contain not only more terms, but a greater wealth of relational information between the terms. Hence it is becoming desirable, and even necessary, to automate at least a part of the development of these knowledge bases. This has been addressed in part by the information retrieval and extraction research communities, primarily through text mining procedures that gather information from existing documents and perform various semantic analyses on the data to determine their fitness to the desired ontology. Examples of this are abundant in the literature, including [1], [2], and [3].

Some ontologies, however, may require the incorporation of knowledge that can best be obtained from the analysis of graphical, not textual, data sources. Ontologies

that represent knowledge about the physical structure of a system are useful in many varied fields such as biology, engineering, and architecture. In each of these domains the concepts need to be organized by their function and/or their spatial relationships to each other. For instance, physicians refer to the combination of the two ventricles and the two atria as the heart. In architecture, the rails, the stiles, and the panes make up a window. The spatial relationships of the parts assist in the determination of their grouping into larger units. Although these relationships could be determined by a domain specialist, the system that is to be represented may be so extensive that it would require a significant amount of the expert's time to construct the ontology using manual methods. Furthermore, it may take a considerable amount of time to determine the spatial relations between the objects on a finer scale than just *is\_a* or *part\_of*. If two dimensional (2D) and/or three dimensional (3D) digital images of these objects are available, it should be possible to automate at least a part of the ontology development process by using image processing and spatial reasoning methods. Although the ontology so created likely still would need to be reviewed by a domain specialist for correctness and completeness, there could be a substantial savings of development time, and perhaps even a more refined and more objective assessment of the spatial relationships between the concepts in the ontology thus created. Given the large number of image sets that exists today compared to the relatively small number of ontologies representing those objects, there clearly is a need for automated algorithms to generate even draft ontologies from image data.

Herein we present an algorithm that addresses precisely those needs; that is, the automated creation of an ontology of physically structured concepts. This algorithm utilizes the Qualitative Spatial Reasoning (QSR) system RCC-8 [4] that is based on the Relational Connection Calculus (RCC) [5]. As a proof of concept, the results produced by our algorithm for collections of 3D anatomical images are compared respectively to an established reference ontology for human anatomy (the Foundational Model of Anatomy [6]) and also to another for amphibian anatomy (AmphibAnat [7]).

## 2 Background and Related Work

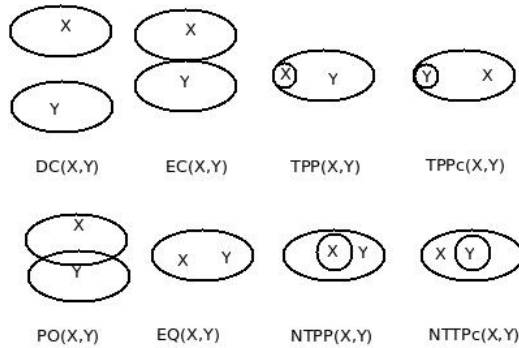
The majority of currently existing ontologies have been assembled laboriously by hand, ideally with domain specialists working closely with experts in ontological representation and reasoning. Some ontologies describe *abstract* concepts, such as the various linguistic ontologies (e.g., [1] and [3]) that attempt to capture the shades of meanings in word usage, and the ontologies that are intended primarily for communication between computers in order to realize the Semantic Web [8]. However, there are also ontologies in use today that are based primarily on real *physical* objects, such as the parts of a machine or the anatomy of an organism (e.g., [7], [9], and [10]). For such sets of data, the primary organizational structure often involves spatial relationships; for example, object A is a part of object B, and object C connects to object D.

Regardless of whether an abstract or a physical knowledge base is being defined, the typical set of events in the construction of an ontology involves the creation of a base set of terms and relationships. It now is becoming more common for this core ontology to be enhanced via various means, usually involving information extraction from various sources such as WordNet [11] for linguistic data, or from a set of publications relevant to the particular domain under consideration, such as in the work being done by Gauch [12] and by Oro [13]. The objective of these efforts is to take terms from an existing core ontology, and then to utilize various data analysis techniques to search for similar terms. The similar terms are scored using some particular set(s) of semantic measures, to provide a possible set of related terms that could be used to enrich the existing base set of ontological terms and relations. These approaches require the prior creation of a core (seed) ontology, and have been limited at this time to obtaining data only from textual sources.

In recent years, the consideration of graphical, not just textual, information for knowledge representation in ontologies has been receiving more attention. One formalization of spatial relationships for the purpose of qualitative reasoning in ontological models is provided by Bittner, Donnelly, Goldberg, and Neuhaus [14]. Their work gives an overview of spatial relationships that are used in the anatomical domain, and discusses how those definitions could be used for automated reasoning. Cohn formalizes a set of relations applied to modeling cell structure [15], explaining how the spatial relationships could be used as part of the manual process of ontology construction. Albath, Leopold, Sabharwal, and Perry [16] proposed RCC-3D, a set of spatial relationships that, like Cohn, are built upon the RCC model for spatial reasoning of [5], but include additional relations to account for the parallel projections that are perpendicular to each of the principal planes in  $R^3$ . For our initial investigation of ontology generation based on spatial relationships, we deemed it sufficient to use the RCC-8 set of relationships proposed by Randell et al [4]. As with Cohn's system, the RCC-8 model was not built specifically with the intention of facilitating automated knowledge base construction, but does provide relationships that can be used for that purpose.

### 3 The RCC-8 Spatial Relationships

A *spatial relationship* has been defined by Blake as “a description of the one or more ways in which the location or shape of a feature is related to the location or shape of another feature” [17]. Spatial relationships are the basis for many problem domains, including fields as diverse as anatomy and Geographic Information Systems (GIS). If one is to do reasoning with spatial relationships, a desirable attribute is to have the set of relations be Jointly Exhaustive and Pairwise Disjoint (JEPD). Simply stated, this means that there should be exactly one relationship that holds between any two objects in the domain — *jointly exhaustive* implying that a relationship exists, and *pairwise disjoint* implying that there is only one applicable relation from the set. This becomes a necessary requirement when the set of relations is to be used for reasoning, as without it the reasoning problem becomes undecidable in many instances and ambiguous in others.



**Fig. 1.** Two-dimensional examples for the eight basic relations of RCC-8 [4]

The JEPD RCC-8 relations for two distinct objects  $X$  and  $Y$  are listed below, and a visual example of each relation in 2D is depicted in Figure 1.

$DC(X, Y)$ :	$X$ is disconnected from $Y$ .
$EC(X, Y)$ :	$X$ is externally connected to $Y$ .
$TPP(X, Y)$ :	$X$ is tangentially a proper part of $Y$ .
$TPPc(X, Y)$ :	$Y$ is tangentially a proper part of $X$ .
$PO(X, Y)$ :	$X$ partially overlaps $Y$ .
$EQ(X, Y)$ :	$X$ is equivalent to $Y$ .
$NTPP(X, Y)$ :	$X$ is nontangentially a proper part of $Y$ .
$NTPPc(X, Y)$ :	$Y$ is nontangentially a proper part of $X$ .

For the above relationships, each has a converse relationship that is also a member of the set. For the RCC-8 set,  $TPP$  and  $TPPc$  are converses of each other, as are  $NTPP$  and  $NTPPc$ . Each of  $EQ$ ,  $EC$ ,  $DC$ , and  $PO$  are converses of themselves; for example, if  $DC(X, Y)$ , then  $DC(Y, X)$  for any two objects  $X$  and  $Y$ . The existence of a converse for each relation means that the algorithms used to discover these relations do not need to consider the order in which object pairs are analyzed.

## 4 The Ontology Construction Algorithm

Unlike most other algorithms for automated generation of ontologies, the ontology construction algorithm presented here does not use a seed ontology, instead building the ontology from scratch. The input for Algorithm 1 is a set of objects (where each object has an associated 3D image) and, for each pair of objects in the collection, the RCC-8 relationship that holds between those two objects. The output of Algorithm 1 is an ontology containing: (a) a node for each of the individual objects, (b) nodes representing groupings of these objects classified using *is\_a* or *part\_of* relationships, and (c) an annotative RCC-8 spatial relationship between each pair of objects.

**Algorithm 1. Ontology Construction**

// Initialize the ontology.

Create the root node *Concept*.

As a child node of *Concept*, add a node for each individual object to be included in the ontology, joined to *Concept* by the relationship *part\_of*.

Create a node *Synonym* as a child node of *Concept* joined by the relationship *is\_a*.

// Handle equivalence relationships.

For each pair of individual (object) nodes X and Y where EQ(X, Y), do:

    Move node Y to be a child of *Synonym* with the relationship *is\_a*.

    Add a relationship stating Y *is\_synonym\_of* X.

End for.

// Handle proper-part relationships.

For each pair of individual (object) nodes X and Y where either TPP(X, Y) or NTPP(X, Y), do:

    If neither X nor Y is an ancestor of the other in the ontology tree:

        Let  $n_1$  be X.

        If X is a descendant of a group node, let  $n_1$  be the most ancestral group node above X.

        Let  $n_2$  be Y.

        If Y is a descendant of a group node, let  $n_2$  be the most ancestral group node above Y.

        If  $n_1$  is not the same as  $n_2$ , move  $n_1$  to be a child of Y, linked by the *part\_of* relationship.

    End if.

End for.

// Handle inverse proper-part relationships.

For each pair of individual objects X and Y where either TPPc(X, Y) or NTPPc(X, Y), do:

    If neither X nor Y is an ancestor of the other in the ontology tree:

        Let  $n_1$  be X.

        If X is a descendant of a group node, let  $n_1$  be the most ancestral group node above X.



```

    Let  $n_2$  be Y.
    If Y is a descendant of a group node, let  $n_2$  be the most ancestral group
    node above Y.
    If  $n_1$  is not the same as  $n_2$ , move  $n_2$  to be a child of X, linked by the
    part_of relationship.
  End if.
End for.
// Handle overlap relationships.
For each pair of individual (object) nodes X and Y where PO(X, Y), do:
  Let  $p_X$  be the parent node of node X and  $p_Y$  the parent node of node Y.
  If X and Y are each direct children of group nodes:
    If  $p_X$  and  $p_Y$  are different nodes
      For each node  $n_c$  that is a child of  $p_Y$ :
        Move  $n_c$  to be a child of  $p_X$ , linked with a part_of relationship.
      End for.
      Remove node  $p_Y$ .
    End if.
  Else if X is a direct descendant of a group node, but Y is not:
    Move Y to be a child of  $p_X$ , linked with a part_of relation.
  Else if Y is directly related to a group node, but X is not:
    Move X to be a child of  $p_Y$ , linked with a part_of relation.
  Else if neither X nor Y is an ancestor of the other:
    Create a new group node as a child of Concept, linked with a part_of
    relationship. Let this node be  $n_0$ .
    Let  $n_1$  be X.
    If X is a descendant of a group node, let  $n_1$  be the most ancestral group
    node above X.
    Let  $n_2$  be Y.
    If Y is a descendant of a group node, let  $n_2$  be the most ancestral group
    node above Y.
    Move  $n_1$  and  $n_2$  each to be children of  $n_0$ , linked with a part_of
    relationship.
  End if.

```

```

End for.
// Handle EC relationships.
For each pair of individual objects X and Y where EC(X, Y), do:
    Repeat the same logic as that for PO(X, Y)
End for.
// Compress group nodes.
Until no more compression is possible, do:
    For each node X in the ontology tree that has exactly one child node Y:
        Let  $p_x$  be the parent of X.
        If X is a group node, move Y to be a child of  $p_x$ , linked with a part_of
        relation, and then remove node X.
        Else if Y is a group node, move each node that is a child of Y to be a
        child of X, linked with a part_of relation, and then remove node Y.
    End for loop.
End until loop.
// Add the RCC-8 relations.
For each pair of objects X and Y, add to the ontology both the RCC-8
relationship between X and Y,  $R(X, Y)$  and the converse of the relation  $R_c(Y, X)$ .

```

The purpose of performing the loop for various relationships separately is that this ensures a more closely associated grouping for objects that are spatially more connected. The EQ relationship results in one of the two terms becoming a synonym of the other. The TPP and NTPP relationships and their inverses ensure that the smaller object is a descendant of the larger object; that is, if X is a proper part of Y, then X will become a descendant of Y. The overlap relationship PO creates the next most closely found relationships in the tree, and the objects related by EC (externally connected) become the third most closely associated objects in the resulting tree. The DC (disconnected) relationship is not included as it does not contribute to any *is\_a* or *part\_of* relationship.

## 5 Comparing Ontology Trees Quantitatively

To objectively evaluate how well the ontology generation algorithm performs, we need to be able to assign a numeric value representing the similarity of a generated ontology to its manually constructed counterpart. There are many algorithms for comparing two trees for similarity, with most such methods attempting to determine whether two trees are isomorphic. For our purposes, complete isomorphism is not the goal. Instead we want to quantitatively measure the similarity of two trees only

considering the leaf nodes of each tree and the overall structure of the two trees (i.e., without regard to the names associated with the internal nodes or the ordering of a node's children). Algorithm 1 systematically labels the internal nodes as groupings are created; thus the names of those nodes likely would not match the names of corresponding nodes in a manually created ontology. Many existing tree comparison algorithms also view the ordering of a node's children as important; however, in an ontology the order of the child nodes has no semantic significance.

Because of these two reasons we developed Algorithm 2 to compare two unordered trees; it does not differentiate the ordering of child nodes, nor does it require matching the names of internal nodes. The algorithm considers all nodes and the *part\_of* relations between the nodes, regardless of whether they are internal nodes (representing a grouping of objects) or leaf nodes (representing individual objects that correspond to the image files). The inputs to the algorithm are two trees (i.e., ontologies), and the output is a real number between 0 and 1 inclusive, where a value of 1 indicates identical trees according to the criteria previously mentioned and a value of 0 represents completely distinct trees according to these same criteria.

It should be noted that this algorithm does not give an absolute score to the similarity of two trees; there is no particular number that is "good enough" to say that two trees are sufficiently similar. Instead the significance of the similarity measure can be explained as follows. Consider a similarity value  $S_A$  determined by comparing a "control" tree to the tree produced by some algorithm A, and a similarity value  $S_B$  determined by comparing the same "control" tree to the tree produced by some other algorithm B. The higher of the two values  $S_A$  and  $S_B$  represents that the corresponding tree is more similar to the control than is the tree that produced the lower comparison value. This in turn indicates that the algorithm associated with the higher similarity value will produce a tree that more closely resembles the target tree. Clearly, an algorithm generating a tree with a higher similarity measure would be more accurate and useful; for our purposes, the generated (ontology) tree likely would require less time and effort for further refinement by a domain specialist.

### **Algorithm 2. Ontology Comparison**

For  $T_1$ , number all internal group nodes from zero for the root to  $n-1$  so that each node has a number.

For  $T_2$ , if a node has a child that is also a node in  $T_1$ , assign that node the same number as its corresponding node in  $T_1$  as long as that number has not otherwise been used in  $T_2$ . Repeat as needed, numbering all possibly corresponding nodes.

For the remaining (non-numbered) nodes in  $T_2$ , number them uniquely beginning with  $n$ .

If there is a path of multiple internal nodes in one tree such that none of these nodes is in the other tree, and such that they have no other children, then collapse this chain of nodes into one single node.

Make a list of all directed paths, giving the start node, the end node, and the number of edges contained in the path in each of the two trees. If the path does not exist in either of the two trees, mark the number of edges with an 'x'.

Let  $D_{\max}$  be the largest number of edges in any path, and  $P$  as the total number of paths in the list, whether from  $T_1$  alone,  $T_2$  alone, or included in both.

For each path  $p_i$ , if it exists in only one of the two trees, score it as zero. If it exists in both trees, the score is  $(D_{\max} - \Delta d)/D_{\max}$  where  $\Delta d$  is the absolute value of the difference in the path length between the two trees for path  $p_i$ .

Sum the scores for all paths  $p_i$  and divide by  $P$  to give the total normalized score for the similarity of  $T_1$  and  $T_2$ .

The calculation of each path score guarantees a value from zero to one; hence the final total score of similarity is always between zero and one. A score of exactly one represents maximum similarity between the trees, and a score of zero represents minimum similarity. Again, these values do not describe the absolute quality of the structures, but only the relative quality; that is, a higher score represents a closer match between the two trees. It is also to be noted that the algorithm is a nondeterministic heuristic, and as such only gives a close score, not an exact value, but one that is sufficiently precise for our purposes. For testing purposes, we were interested in comparing a “control” tree (or manually created ontology) with the ontology generated by our algorithm; thus as our algorithm evolved we could then say whether the changes were improvements or not.

## 6 Examples of Ontology Generation from 3D Images

For testing purposes it was difficult to find a nontrivial ontology for which an associated set of 3D images was also available; however, as proof of concept for our algorithms, we obtained a set of one hundred 3D image data files (in OBJ format) from Zygote Media Group, Inc. [18] that collectively represent objects that comprise the human brain. These particular data files were selected in part because they correspond to a section of the Foundational Model of Anatomy (FMA) ontology [19]; thus giving us the opportunity to compare the output of our algorithm to an established, mature ontology.

To determine the relationship between each pair of objects, we used a program developed by Albath, Leopold, Sabharwal, and Perry [16] that takes as input two data files (in OBJ format) and determines which one of the RCC-8 relationships holds for the two objects. There were  $(100! / (2! \times 98!)) = 4950$  pairs of objects in the human brain data collection, and thus 4950 RCC-8 relationships for the entire set of objects. The computed RCC-8 relations were then manually verified by examining the images using a program developed for viewing 3D reconstructions [20]. An implementation of Algorithm 1 was subsequently given as input the collection of 3D images and the RCC-8 relation determined for each pair of objects. The generated ontology was then compared to a portion of the FMA that includes the corresponding parts of the human brain. Since Algorithm 1 has no provision for supplying precise, meaningful names for the group nodes that are generated during processing, it should be noted that an expert likely would be required to rename those nodes.

Brain	Concept_brain
Forebrain	Group_1
Diencephalon	Group_2
Telencephalon_or_cerebrum	Group_3
Gray_matter_structure_of_cerebral_hemisphere	Group_4
Basal_ganglia	Telencephalon_or_cerebrum
Globus_pallidus	Globus_pallidus_lateral
Globus_pallidus_lateral	Globus_pallidus_medial
Globus_pallidus_medial	Putamen
Striatum	Substantia_nigra
Caudate_nucleus	Subthalamic_nucleus
Putamen	Group_5
Substantia_nigra	Group_6
Subthalamic_nucleus	Caudate_nucleus

**Fig. 2.** On the left is a portion of the FMA ontology for the brain. On the right is a section of the ontology generated by Algorithm 1. Each shows only the *part\_of* relationships.

For brevity, only a portion of the generated ontology and the corresponding section of the FMA ontology are shown in Figure 2. Applying the calculations of Algorithm 2 to the FMA ontology and the results produced by Algorithm 1 for this dataset, we found that the maximum path length ( $D_{\max}$ ) was 8 and that there were 348 total paths. Of these, 36 scored 1.000, 35 scored 0.875, 26 scored 0.750, 14 scored 0.625, 8 scored 0.500, and the remaining paths had a score of zero. Completing the calculations gives an overall similarity score of 0.284 for the comparison.

For a second test, a set of thirty-one images representing the bones of the head of a frog was obtained from MorphologyNet [20]. The results were compared to AmphibAnat [7], which is an established ontology containing multi-species anatomical data for amphibians. To create the ontology using these files, identical processing was performed to that previously described for the human brain files.

Analysis of the results for this set of object files produced an overall similarity score of 0.276 for the comparison between the AmphibAnat ontology and the ontology produced by Algorithm 1 for this data. In part, this may suggest that similar logic was used by both the FMA and AmphibAnat groups when manually creating their respective ontologies.

Although the overall similarity scores were not as close to 1 (a perfect match) as one would hope, there still are three significant advantages to using Algorithm 1 to construct a draft ontology: (1) each leaf node term in the ontology has an image file associated with it, thereby providing a visual reference for the object, (2) the resulting ontology is not simply a partology, but also contains richer semantic content by virtue of the RCC-8 relations that have been assigned between leaf node terms, and (3) the ontology thus created can be edited more quickly and easily than would be the case if the developer was starting completely from scratch. Although it is hoped that subsequent algorithms can achieve higher scores, it is doubtful that a perfect similarity can be obtained since there is no one “right” ontology for any set of data.

## 7 Future Work

In the future we plan to consider additional spatial ordering relationships which capture the orientation of the parts in relation to each other (e.g., relations such as *dorsal*, *ventral*, *rostral*, and *proximal*). Incorporating such relations into the construction process could further aid in the accuracy and completeness of the ontology generated by the algorithm.

We also note that the RCC-3D relationships of [16] have refinements that consider perspective. Although this was not deemed necessary for consideration in our preliminary work, these refinements will be examined for future inclusion in the reasoning process to determine if they are in any way beneficial to the ontology construction process and the resulting representation of knowledge.

Another aspect of this approach that we intend to investigate in more detail is the generation of the extraneous *group* nodes that can occur. Although the construction algorithm collapses these nodes under certain circumstances, and although the similarity calculations account for this discrepancy, further work needs to be done in this respect. We hope to examine more diverse datasets and corresponding established ontologies in an attempt to determine modifications to the generation algorithm that will perform term grouping in a manner more consistent with that performed by the domain expert. This should result in a generated ontology that more closely resembles the manually created counterpart, and thus would result in an ontology that would require less refinement by the domain expert after its initial automated construction.

## 8 Summary and Conclusions

Because of the vast amount of time required for manual creation of ontologies, algorithms for generating and extending ontologies using data from multimedia resources are needed. Herein we presented an algorithm that utilizes qualitative spatial reasoning over 3D images to automate the initial construction of an ontology.

Using a simple method for quantitatively measuring the similarity of two trees, as proof of concept we compared the results obtained by the ontology generation algorithm for two different collections of anatomical 3D images to established, manually constructed reference ontologies for those same objects. Although the generated ontologies for these examples were not as close a match to the target ontologies as one would hope, there still are significant advantages to using this method; namely: (a) each leaf node term in the ontology will have a visual reference associated with it, (b) the resulting ontology will contain richer semantic content than simply a partology by virtue of the RCC-8 relations that will have been assigned between leaf node terms, and (c) the resulting ontology likely will be more easily edited than would be the case if the developer was starting completely from scratch.

The primary advantage of computer automation is a savings in the time needed to accomplish a task. It is hoped that the approach presented in this paper will facilitate the development of comprehensive ontologies that represent knowledge about the physical structure of a system.

**Acknowledgments.** This work was supported by NSF under awards DBI-0445752 and DBI-0640053.

## References

1. Luong, H., Gauch, S., Speretta, M.: Enriching Concept Descriptions in an Amphibian Ontology with Vocabulary Extracted from WordNet. In review (2010)
2. Kang, S., Lee, J.: Semi-Automatic Practical Ontology Construction by Using a Thesaurus, Computational Dictionaries, and Large Corpora. In: Annual Meeting of the ACL, Proceedings of the Workshop on Human Language Technology and Knowledge Management (2001)
3. Yang, J., Wang, L., Zhang, S., Sui, X., Zhang, N., Xu, Z.: Building Domain Ontology Based on Web Data and Generic Ontology. In: Proceedings of the IEEE/WIC/ACM International Conference on Web Intelligence (2004)
4. Randell, D.A., Cui, Z., Cohn, A., Nebel, B., Rich, C., Swartout, W.: A Spatial Logic Based on Regions and Connection. In: Proceedings of the Third International Conference on Principles of Knowledge Representation and Reasoning, KR 1992, vol. 1992, pp. 165–176 (1992)
5. Cohn, A., Bennett, B., Gooday, J., Gotts, N.: RCC: A Calculus for Region Based Qualitative Spatial Reasoning. *GeoInformatica I* 1997, 275–316 (1997)
6. Rosse, C., Mejino, J.: A Reference Ontology for Biomedical Informatics: The Foundational Model of Anatomy. *Journal of Biomedical Information* 36, 478–500 (2003)
7. Maglia, A., Leopold, J., Pugener, L., Gauch, S.: An Anatomical Ontology for Amphibians. In: Pacific Symposium on Biocomputing, vol. 12, pp. 367–378 (2007)
8. Shadbolt, N., Hall, W., Berners-Lee, T.: The Semantic Web Revisited. *IEEE Intelligent Systems* 2006, 96–101 (2006)
9. Baldock, R., Bard, J., Kaufman, M., Davidson, D.: A Real Mouse for Your Computer. *BioEssays* 14, 501–502 (1992)
10. Sprague, J., et al.: The Zebrafish Information Network: The Zebrafish Model Organism Database. *Nucleic Acids Research* 34, D581–D585 (2006)
11. Miller, G.: WordNet: A Lexical Database for English. *Communications of the ACM* 38, 39–41 (1995)
12. Speretta, M., Gauch, S.: Using Text Mining to Enrich the Vocabulary of Domain Ontologies. *Web Intelligence* 2008, 549–552 (2008)
13. Oro, E., Ruffolo, M., Sacca, D.: Ontology-Based Information Extraction From PDF Documents With XOnto. *International Journal on Artificial Intelligence Tools* 18, 673–695 (2009)
14. Bittner, T., Donnelly, M., Goldberg, L., Neuhaus, F.: Modeling Principles and Methodologies – Spatial Representation and Reasoning. *Anatomy Ontologies for Bioinformatics: Principles and Practice* 2008, 307–326 (2008)
15. Cohn, A.: Formalising Bio-Spatial Knowledge. In: FOIS 2001, vol. 2001, pp. 198–209 (2001)
16. Albath, J., Leopold, J., Sabharwal, C., Perry, K.: Efficient Reasoning with RCC-3D. In: Proceedings of the 4th International Conference on Knowledge, Science, Engineering, and Management, Belfast, Ireland (2010)
17. Blake, L.: Spatial Relationships in GIS – An Introduction. *OSGeo Journal* 1, 1–3 (2007)
18. Zygote Media Group, Inc.: (2010), <http://www.zygote.com/>
19. Foundational Model of Anatomy (2010), <http://sig.biostr.washington.edu/projects/fm/>
20. Leopold, J., Maglia, A.: Evaluation of MorphologyNet as a Virtual Dissection Experience. In: Proceedings of World Conference on Educational Multimedia, Hypermedia, and Telecommunications, vol. 2006, pp. 2848–2851 (2006)

# Facilitating Experience Reuse: Towards a Task-Based Approach

Ying Du<sup>1</sup>, Liming Chen<sup>2</sup>, Bo Hu<sup>1</sup>, David Patterson<sup>2</sup>, and Hui Wang<sup>2</sup>

<sup>1</sup> SAP Research CEC Belfast, UK

{ying.du,bo01.hu}@sap.com

<sup>2</sup> University of Ulster, UK

{l.chen,wd.patterson,h.wang}@ulster.ac.uk

**Abstract.** This paper proposes a task-based approach to facilitate experience reuse in knowledge-intensive work environments, such as the domain of Technical Support. We first present a real-world motivating scenario, product technical support in a global IT enterprise, by studying of which key characteristics of the application domain and user requirements are drawn and analysed. We then develop the associated architecture for enabling the work experience reuse process to address the issues identified from the motivating scenario. Central to the approach is the task ontology that seamlessly integrates different components of the architecture. Work experience reuse amounts to the discovery and retrieval of task instances. In order to compare task instances, we introduce the dynamic weighted task similarity measure that is able to tuning similarity value against the dynamically changing task contextual information. A case study has been carried out to evaluate the proposed approach.

**Keywords:** Task Management, Knowledge Management, Semantic Similarity Measure.

## 1 Introduction

Experience is the specific knowledge resided in a problem solving process and context [1] [2]. It is usually acquired through repeatedly performing work activities and executing daily tasks within a domain. Organisations can benefit from work experience reuse in terms of improved productivities and cost reduction.

Previous studies [3] [4] [5] have identified the relationship between task and work experience. Work experience is incrementally gained from the execution of tasks, and successful task execution certainly requires the possession of relevant work experiences. In order to formally capture the relationship, we introduce the notion of task model as the carrier of work experience and propose a task-based approach to experience modelling, representation, reasoning and reuse.

In the proposed approach, work experience is captured in a task ontology that comprises the entities involved in the task execution, such as documents, emails, and people. The relationships between these entities and the task are explicitly represented by the properties of the task ontology. This provides rich knowledge structures for



describing how a piece of work is organised, structured and accomplished. Work experience is formally represented as task instances. As such they can be reused by other colleagues who have similar tasks at hand.

The paper is organised as follows. We first give an overview of the SAP technical support team in Section 2 and highlight the user requirements of experience reuse. Section 3 introduces the task-based approach and an architecture enabling the reuse of work experience in the technical support application domain. Section 4 describes algorithms for computing task similarity as a means of experience retrieval. The proposed approach and algorithms are evaluated in Section 5. We discuss related work in Section 6 and conclude the paper in Section 7.

## 2 Motivating Scenarios

Our task-based experience reuse approach is built on a real world use case – the SAP Business ByDesign (ByD) Technical Support Team. The team is responsible for providing technical support to ByD customers around the globe. The general process of the technical support is that customers use IT tickets to describe technical problems and submit these tickets to technical support technicians for requesting solutions.

To identify key characteristics of experience reuse, we carried out a one-week on-site case study with the SAP ByD support team in Ireland. The finding shows that experience reuse occurs frequently. When receiving a new IT Ticket, technicians constantly make reference to existing ones. This is because similar tickets provide hints as to which information resources can be used for solving the new problem. In addition, technicians frequently communicate with their colleagues thus to find the right people to collaborate with. Based on the observations, the problem-solving process for a given IT ticket consists of two major tasks: search similar IT Tickets, and seeking the ‘right’ person to collaborate with. In the following, we will discuss these two work aspects in detail.

Technicians manage their daily tasks in an IT Ticket Management System. The system employs a keyword-matching mechanism to support searching similar tickets. Keywords are manually selected from the description of a ticket. The drawbacks of the keyword-matching mechanism are evident: First, as most tickets contain complicated technical issues, the manually selected keywords may not be able to precisely reflect the technical issues. Second, the keyword-matching mechanism oversimplifies the ticket comparison problem, as in some cases, similarities are determined by other factors rather than the manually defined keywords.

According to our analysis, around 90% tickets are solved collaboratively within a group of people. The IT ticket management system is designed for enabling tickets tracking rather than supporting work collaboration. Typically, technicians have to leverage their work experience to identify the ‘right’ person to collaborate with.

To solve the above problems, we will follow the following two paths. First, the keyword-matching mechanism will be replaced by the dynamic weighted task similarity algorithms, and ticket similarities will be calculated by leveraging the advantage of

semantic technologies. Second, work collaborations will be captured and represented in terms of the task model ontology, and will be fully supported in our proposed approach.

### 3 A Task-Based Approach to Experience Reuse

#### 3.1 Task-Based Experience Modelling

A task can be viewed as a number of activities associated with a particular goal. An activity usually involves a number of entities. An entity refers to any object involved in task execution, such as digital documents, participants, and other resources. The semantic model of task can capture the involved entities, the relations among these entities, and the relations between entities and tasks (see Fig. 1). For example, the entity *Collaborator* describes the persons who participate in task execution; the *Email* entity records email communications between collaborators. We formally conceptualise the involved entities into the Task Model Ontology (TMO) [6].

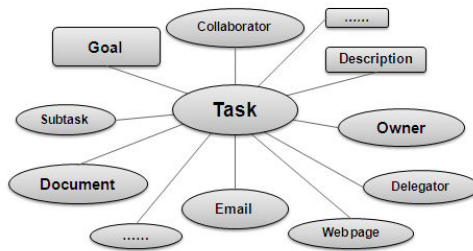


Fig. 1. A conceptual task model

Entities captured in the task model describe how a task is performed. The task model is to concretise the work experience but it is not an entire reflection of work experience. This is because work experience is not easy to be completely captured. In order to successfully accomplish a task, the person must know how to apply the associated resources and how to solve the problems appearing in the task. Therefore, the task model can approximate the work experience that one accumulates from the task execution. Hence, the task model can be regarded as a carrier of work experience.

#### 3.2 The Architecture for Task-Based Experience Reuse

Fig.2 shows the system architecture supporting the proposed task-based experience reuse in the technical support application domain. The architecture enables two processes. The first one is the semantic annotation process in which a ticket is automatically annotated with metadata and a task instance is generated from a ticket. The second one is the experience reuse process based on recommending similar tasks.

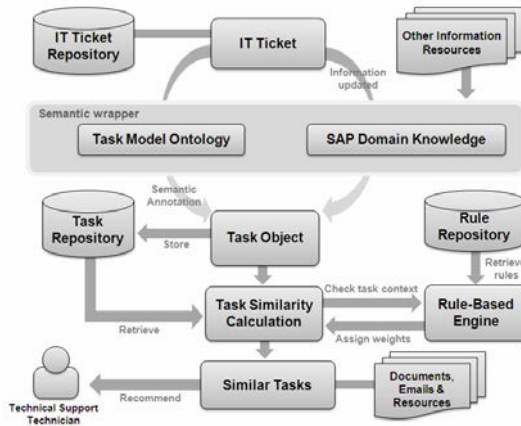


Fig. 2. Task-driven work experience reuse in technical support

### 3.2.1 The Semantic Annotation Process

The Semantic Annotation process integrates tickets with other useful information resources (such as customer information and system information), adds semantic metadata, and transfers these semantic enriched information items into a task object.

A ticket consists of both structured data and unstructured data. Structured data includes information about the customer (ticket creator), involved technicians, time stamps and other structured information items. Unstructured data contains the problems described by a customer, and the solutions discussed and provided by technicians. The Semantic Wrapper component is used to wrap both the structured and unstructured data into a task instance generated from the TMO task ontology.

The TMO is an upper-level task ontology which captures general aspects of a task, such as task attributes like delegator, owner, collaborator and status. The structured data of tickets can be directly reused to generate task instances. For example, the customer who creates the ticket is described as task delegator in the task ontology. The TMO treats a customer as an object rather than a piece of information. It uses a class, e.g., *person* to model and represent customer information by a number of attributes, such as *hasLocation*, *hasContactNumber*, *hasPersonalProfile*. As such, the TMO provides an ontological model for representing a ticket.

As an upper-level task ontology, the TMO is not able to capture the characteristics of a particular domain. Therefore, we introduce the concept of *domainTerm* as a new task attribute to capture relevant domain knowledge. Here, *domainTerm* refers to domain-specific terms that are commonly understood in a particular application domain. These terms are embedded in ticket descriptions and have to be explicitly described in task instances. Therefore, we develop the SAP Domain Vocabulary component to filter out domain terms. For example, if the term “CRM” is described in a ticket description and it has been recorded in the SAP Domain Vocabulary component, they will be automatically added as domain knowledge terms to the corresponding task in the semantic annotation process.

### 3.2.2 The Experience Reuse Process

The first step of reusing past work experiences is to identify which experiences are relevant to the current work at hand. In our proposed approach, technician’s current work, i.e., technical problems to solve, is represented in a task instance. When the solution is provided, relevant work experiences are captured in the task instance. Hence, the task instance becomes a carrier of work experience, and experience reuse can be achieved by comparing task similarities.

This process is triggered when a user searches for similar tasks to a particular task at hand. The search results will be a list of similar tasks with different rankings. The similarity ranking between two tasks will change over time when a task is updated with new information that will affect the similarity calculation.

In the calculation process, the following two factors have impact on task similarities: the similarity of task attributes and the human defined rules. Similarity of task attributes can be calculated based on string comparison. Human defined rules are the representation of human knowledge in which technicians can describe their expertise and work experience into rules to affect task similarities. Human generated rules will be invoked by Rule-Based Engine to adjust task similarity according to the dynamic weighted task similarity algorithms defined in Section 4.

## 4 Task Similarity Measurement

### 4.1 The Dynamic Weighted Task Similarity Measurement

Motivated by the Tversky’s feature-theoretical approaches (the contrast model [7] and the ratio model [8]), we define a global function as shown in Equation 1 to calculate task similarities. Let  $T_a$  and  $T_b$  denote two tasks,  $\omega_{Att_k}$  denote the weights for feature  $k$ . The value of  $Sim(T_a, T_b)$  is between 0 and 1. The value 0 means  $T_a$  is not similar to  $T_b$ , whereas any value greater than 0 indicates  $T_a$  and  $T_b$  are to some extent similar. The greater the value is, the more similar the two tasks are.

$$Sim(T_a, T_b) = \sum_{i=1}^n \omega_{Att_k} * Sim_{Att_k}(T_a, T_b) \tag{1}$$

The similarity function  $Sim_{Att_k}(T_a, T_b)$  computes the similarity of each task attribute (from  $Att_1$  to  $Att_k$ ).  $n$  is the number of compared task attributes. In the computation process, we use the Jaccard Similarity [9] to calculate the string similarity of task attributes. For example, if the task name (attribute) of  $T_a$  and  $T_b$  is literally equivalent, the similarity value  $Sim_{Name}(T_a, T_b)$  equals 1. On the contrary, the value is 0 when no common words are shared in the attributes.

$Sim_{Att_1}(T_a, T_b) \dots Sim_{Att_2}(T_a, T_b) \dots Sim_{Att_k}(T_a, T_b)$  return a number of task attribute similarity values in the range between 0 and 1. Depending on user defined rules, the similarity of individual attributes are assigned with different weights: from  $\omega_{Att_1} \dots \omega_{Att_2} \dots$  to  $\omega_{Att_k}$ . The weights (Equation 2) represent the dynamically changing context and integrate this into our similarity algorithms. It also accommodates our finding that different task attribute has different importance during similarity calculation.

$$\omega_{Att_k} = \frac{\frac{1}{n} + \gamma_k}{\sum_{i=1}^n \frac{1}{n} + \gamma_k} \quad (2)$$

Here,  $n$  is the number of compared task attributes. All attributes will be assigned to the weight of  $1/n$  if no user defined rules can be applied to the similarity computation process. In this case, all attributes share equivalent importance and hence  $\omega_{1...} = \omega_{Att_k} = 1/n$  and  $\gamma_k = 0$ .  $\gamma_k$  is a parameter defined in user defined rules with the aims of adjusting task similarities according to human knowledge.

## 4.2 Use Domain Knowledge for Similarity Computation

We propose the concept of Task Similarity Rule (TSR) to formally model and represent domain knowledge for task similarity calculation. Technicians can specify their own views on task similarity in TSR and thus improving similarity calculation result based on their domain expertise. This is achieved by defining a number of conditions in a TSR. Each condition represents a task context which consists of a task attribute with predefined value. Features of the world become context through their use [10] [11]. By extending this idea, features of a task become task context through their use in similarity computation. Features are modelled by the task attributes in the task model ontology.

```

IF Ta.delegator = Tb.delegator   AND
   Ta.domainTerm  $\supseteq$  {'v 1.2', 'CRM'}   AND
   Tb.domainTerm  $\supseteq$  {'v 3.1', 'CRM'}   AND
THEN
 $\gamma_1 = \gamma_2 = \dots = \gamma_k = -\frac{1}{n}$ 

```

**Fig. 3.** A TSR refines the weight of Ta and Tb attributes

Fig.3 shows a TSR example. The task attributes: *delegator* and *domainTerm* are associated with particular values, denoting the task context information in the TSR. This TSR states that *Ta* and *Tb* are not similar in such task contexts. It expresses the domain knowledge that the system problems about *CRM* are different in system (version) *v1.2* and system *v3.1*. Hence, values of  $\gamma_1 = \dots = \gamma_k$  are assigned to  $-1/n$ . According to the Equation 2,  $\omega_{Att_1 = \dots = Att_k} = 0$  and thus the function  $\text{Sim}(Ta, Tb)$  is equal to 0. Without considering this human knowledge presented in the TSR, *Ta* and *Tb* will be treated as similar tasks given that they have common attribute values, for example, they share the same *delegator* and *domainTerm*.

In addition to determining dissimilarity between tasks, a TSR also can define that two tasks are absolutely similar to each other in which the  $\text{Sim}(Ta, Tb)$  must be equal to 1. In this case, task similarity is solely depended on the TSR. For example, a technician has the knowledge that *Ta* and *Tb* must be similar tasks if *Ta* and *Tb* are from the same delegator and their domain terms include 'v1.2', 'CRM', and 'XH error'. To capture the knowledge, we can define the value of  $\gamma_k$  to let  $\text{Sim}(Ta, Tb) = 1$ . Fig. 4 shows an example of a TSR defining the absolute similarity between two tasks.

$$\begin{aligned}
 &\text{IF } Ta_{\text{delegator}} = Tb_{\text{delegator}} \quad \text{AND} \\
 &\quad Ta_{\text{domainTerm}} \cap Tb_{\text{domainTerm}} \supseteq \{ 'v 1.2', 'CRM', 'XH error' \} \quad \text{AND} \\
 &\text{THEN} \\
 &\quad \gamma_k = \frac{1 - \text{Sim}_{\text{Att}k}(Ta, Tb)}{n * \text{Sim}_{\text{Att}k}(Ta, Tb) - 1}
 \end{aligned}$$

**Fig. 4.** A TSR defines the absolute similarity of Ta and Tb

In some situations, domain knowledge is not sufficient to support such determination. For example, a technician has the domain knowledge that the tasks associated with *domainTerm* ‘v 1.2’ and ‘CRM’ should have similarity and especially the tasks delegated from the same customer are more similar. In this case, the attribute *delegator* has great impact on task similarity. Therefore, the weight value of *delegator*:  $\omega_{\text{delegator}}$  should be increased by assigning a value to  $\gamma_{\text{delegator}}$ . According to the above knowledge, a TSR can be generated as shown in Fig.5.

$$\begin{aligned}
 &\text{IF} \\
 &\quad \text{CCI: } [ Ta_{\text{delegator}} = Tb_{\text{delegator}} \quad \text{AND} \\
 &\quad \quad Ta_{\text{domainTerm}} \cap Tb_{\text{domainTerm}} \supseteq \{ 'v 1.2', 'CRM' \} ] \\
 &\quad \text{OCI: } [ (Ta_{\text{timeStamp}}, Tb_{\text{timeStamp}}) \rightarrow \text{Closer; Similarity} \rightarrow \text{Increase} ] \\
 &\text{THEN} \\
 &\quad \gamma_{\text{Delegator}} = \frac{\theta}{\varphi}
 \end{aligned}$$

**Fig. 5.** A TSR increases task similarity by defining compulsory/optional -context

The Compulsory Context Information (CCI) is the task context required to ensure a TSR is applicable in similarity computation. Task similarity will be determined by the string-comparison of task attributes when CCIs are not fulfilled. In this case, all task attributes share the same importance during similarity measure which means their weights ( $\omega_{\text{Att}k}$ ) are equal to  $1/n$  and hence the value of  $\gamma_k$  is equal to 0.

The Optional Context Information (OCI) will not affect the applicability of a TSR. The role of OCI is to improve the accuracy of similarity measure. Based on the task context defined in an OCI, the similarity will be further adjusted (increase or decrease) by changing the value of  $\gamma_k$ .

We will use the following example to demonstrate how to calculate the values of  $\theta$  and  $\varphi$  to get  $\gamma_{\text{delegator}}$ . Suppose that three tasks (Tb, Tc and Te) are meet the CCI defined in Fig.5, and they are the similar tasks to Ta. Their similarities to Ta will be calculated by the function  $\text{Sim}(Ta, Tc)$ ,  $\text{Sim}(Ta, Td)$  and  $\text{Sim}(Ta, Tn)$ . The OCI is used for further adjusting the similarity value of each function. In the three tasks, Tn’s creation data (*timeStamp*) is the closest one to Ta, followed by Tc and then Td.  $\varphi$  is the total number of the tasks that meet all the CCI. Here, these tasks will be Tc, Td, and Tn and therefore  $\varphi = 3$ .  $\theta$  is the task-ranking value based on the OCI. The highest ranked task (Tn) will be assigned to the value of  $\varphi$  (in this case  $\theta = \varphi$ ), the second

highest ranked task ( $T_c$ ) will be assigned to the value of  $\varphi-1$ , and the third highest ranked task ( $T_d$ ) will be assigned to the value of  $\varphi-2$ . The highest ranked task is  $T_e$  because  $T_e$ 's creation date is closest to  $T_a$ . Therefore, in function  $\text{Sim}(T_a, T_n)$ , its  $\gamma_{\text{delegator}} = 3/3 = 1$ ; in  $\text{Sim}(T_a, T_c)$ ,  $\gamma_{\text{delegator}} = 2/3$ ; in  $\text{Sim}(T_a, T_d)$ ,  $\gamma_{\text{delegator}} = 1/3$ . In the above example, the values of different  $\gamma_{\text{delegator}}$  are adjusted according to the context defined in the OCI.

## 5 Applying the Task-Based Approach

### 5.1 Semantic Task Generation and Task Similarity Calculation

As mentioned in Section 3.2.1, an IT Ticket consists of both structured and unstructured data. The structured data, such as *ticket name*, *creation date-time*, (submitted by) *customer company*, and *ticket processor*, will be directly transferred to a task instance. The unstructured data is included in ticket descriptions. We developed the *SAP Domain Vocabulary* component to extract domain terms from the unstructured data of IT Tickets. The component contains a number of pre-defined SAP domain terms used in ByD systems. These domain terms, e.g. 'CRM', 'FIN', and 'server error', are manually defined by the ByD domain experts. When a task instance is automatically generated from a ticket, the component takes the ticket description as input and automatically checks if any domain term is included. If domain terms are found, they will be added to the task instance. The task attribute *domainTerm* holds the value of these found domain terms.

To calculate task similarity, we have to determine which task attributes have impact on similarity measurement. By analyzing the characteristic of the motivating scenarios, we use the following five task-attributes: *name*, *domainTerm*, *delegator*, *owner* and *collaborator* in task similarity calculation.

Here is an example illustrating the process of task similarity computation. Suppose tickets,  $T_i\text{-a}$ ,  $T_i\text{-b}$  and  $T_i\text{-c}$ , are wrapped in three task instances T1, T2 and T3.  $\text{Sim}(T1, T2)$  denotes the similarity between T1 and T2, and  $\text{Sim}(T1, T3)$  denotes the similarity between T1 and T3. To protect customer privacy, we replace sensitive contents, e.g., customer name, support technician name, etc., with dummy data as shown in Table 1.

Once we decide the attributes used for task similarity computation, we can calculate similarity for each attribute, e.g.,  $\text{Sim}_{\text{Attk}}(T1, T2)$  representing the similarity between task T1 and T2 on the attribute *Attk*. Attribute similarity is calculated using the Jaccard similarity algorithm that is based on string comparison. For each task attribute, string similarity is displayed in Table 2.

If no user-defined rule is applicable to the tasks, the value of  $\gamma_k$  (Equation 2) is equal to 0. Then, all 5 weights (from  $\omega_{\text{name}}$  to  $\omega_{\text{collaborator}}$ ) are equal to 1/5. This means each task attribute shares the same importance during the similarity calculation. According to Equation 1,  $\text{Sim}(T1, T2) = 0.28$ , and  $\text{Sim}(T1, T3) = 0.33333334$ . When each task attribute has the same weight, we refer to the resulting similarity as the 'Average weighted' task similarity.  $\text{Sim}(T1, T3)$  is greater than  $\text{Sim}(T1, T2)$  and which means T3 is more similar to T1.

**Table 1.** Task attributes and attribute value

Attributes and values	
T1	{ name: CompanyX- CRM error; domainTerm: CRM, server error, v1.2; delegator: c10203; owner: e89059; collaborator: ; timeStamp: 21.11.2009 19:40:08}
T2	{name: Exception; domainTerm: CRM, exception, v1.2; delegator: c10203; owner: e16832; collaborator: e78653; timeStamp: 21.11.2009 15:28:01}
T3	{name: CRM error; domainTerm: CRM, server error, v1.2; delegator: c35029; owner: e73847; collaborator: e73847; timeStamp: 27.09.2009 17:27:50}

**Table 2.** String similarities between each task attribute

Attribute	SimAttk(T1,T2)	SimAttk(T1,T3)
name	0.0	0.6666667
domainTerm	0.4	1.0
delegator	1.0	0.0
owner	0.0	0.0
collaborator	0.0	0.0

Three TSRs are defined in the rule repository and the rule engine finds that the rule (in Fig.4) is applicable to Sim(T1, T2). By applying the rule, the value of  $\gamma_{delegator}$  is equal to 1 and hence  $\omega_{delegator}=3/5$ . Accordingly, the weights for other task attributes are equal to: 1/10. The importance of delegator is increased. The weights are affected by the rule, and we refer to this approach as ‘Dynamic weighted’ task similarity calculation. Comparing to the ‘Average weighted’ approach, the similarity between T1 and T2 is increased after applying the rule. Sim(T1, T2) is equal to 0.64. Since no rule is applicable to Sim(T1, T3), the value is still equal to 0.33333334. The result shows that T2 is more similar to T1 than T3.

**5.2 Evaluation**

We used the human judgements of similarity as the benchmark in the evaluation of the proposed task similarity algorithms. The similarity between the tickets (with ID: I3284, I3099 and I3270) has been confirmed by four ByD support technicians who are the solution providers of I3284. According to their expertise, solutions of I3099 and I3270 can be reused to solve the problem reported in I3284. We will use this as the benchmark to carry out the following three experiments. The first experiment is to use the keyword-search function in the ByD ticket management system to find the most similar tickets to I0284. The second experiment is to apply the average weighted similarity algorithms to calculate task similarities. The third experiment is to define a number of TSRs and apply the dynamic weighted similarity algorithms.

In the first experiment, we have to identify the keywords to similar tickets searching. We ask the four technicians to provide one term that they will use as the keyword to search similar IT Tickets to the ticket I3284. Based on their recommended keyword, the search engine returns one hundred and thirty one tickets ranked from number 1 (the most similar) to number 131 (the most dissimilar). The ticket I3735 it is the most similar ticket to the searched keyword. The ranking value assigned to I3284 is 75, to I3099 is 59, and to I3270 is 103.



In order to carry out the task-based similarity measurement, tickets have to be transferred to semantic task instances. Therefore, we create an experiment dataset which contains 221 task instances that are automatically created from the selected tickets in the ByD ticket system. The task instance created from ticket I3284 will be compared with other task instances by using the average weighted task similarity algorithms and the dynamic weighted task similarity algorithms.

Three TSRs are developed under the collaboration with the four technicians. These TSRs capture the knowledge that the technicians used for judging ticket similarities.

**Table 3.** Similarity ranking: comparing ticket I3284 with other top-ranked tickets

	I3735	I3608	I3461	I3445	I3191	I3099	I3180	I3280	I3151	I3270
Keyword search	1	2	3	15	25	59	60	82	96	103
Average weighted	2	5	6	9	10	3	4	1	8	7
Dynamic weighted	DS	DS	DS	3	6	2	4	1	5	1

Table 3 shows the ranked similarity results between ticket I3284 and other ten tickets. Similarity values in the columns of ‘Average weighted (AW)’ and ‘Dynamic Weighted (DW)’ are calculated based on task instances that are created from the corresponding tickets. As mentioned above, I3735 is regarded as the most similar ticket under keyword-based mechanism and it gets the highest ranking value in the ‘Keyword-search’ column. The AW also assigns a high ranking value (the 2<sup>nd</sup> similar ticket) to I3735. This is because the AW similarity algorithms rely on string comparison. However, I3735 is ranked as dissimilar (marked as DS in Table 3) in the DW ranking. The dissimilarity is assigned by a TSR defined by the technicians.

According to the benchmark, I3270 and I3099 should be assigned with the highest similarity ranking value. By comparing the ranking values calculated by the three approaches, the DW ranking is the most similar one to the human judgements in terms of similarity measurement.

## 6 Related Work

The Experience Factory [12] defines a logical and physical infrastructure to support software companies in continuous learning from previous experience of product development. [13] extends the Experience Factory to integrate AI technologies for manipulating experiences. These studies follow a top-down approach for experience reuse, i.e., managing work experience at the organisational level. Task-driven approaches stress the need of populating experience reuse in a bottom-up fashion.

[14] [15] propose the pattern based approaches to managing and reuse valuable work experience. The concept of pattern (e.g., task pattern and activity pattern) is introduced as a means of capturing work experience. A *pattern* is the abstraction of a group of similar tasks and it can be used as a template for creating new tasks. The pattern based approaches require the extensive participation of users and their contribution in pattern creation and maintenance. This brings the difficulties to realise the approaches in real-world scenarios.

Similar to our work, [16] argues that previous similar tasks can provide guidance for current task execution. Task similarity is calculated by a document classification system [17]. EPOS [18][19] develop algorithms to calculate task similarity in terms of task context. Different weights are assigned to possible relations (e.g., ‘is-a’ and ‘part-of’) between concepts. However, the methods for defining these weights are not introduced in their work. For task similarity measurement, we argue that different task attributes should have different impacts on similarity measure in different task context. The impacts are represented by the dynamically assigned weights during similarity calculation.

## 7 Conclusion

In this paper, we introduce a semantic-enabled task-based approach to facilitating work experience reuse. The idea is motivated by the case study with the SAP Business ByDesign technical support team in which work experience reuse is demanded by the technical support technicians. We have extended the upper-level TMO with relevant domain knowledge and leveraged the enhanced TMO as the carrier of work experience. After developing the task ontology to capture and represent work experience, we have focused on the issues of experience retrieval. To this end, we have conceived and developed the dynamic weighted algorithms to compute task similarities. Once task similarities are correctly calculated, the work experience modelled in the task instance can be retrieval and reused. The proposed approach has been evaluated with real-world data from which findings and conclusions have been drawn.

This study highlights several open issues for future work. Currently, the similarity measure is calculated based on the string comparison using the Jaccard similarity. Leveraging semantic relations between task instances will be the potential research direction for improving the accuracy of similarity measurement. Another interesting issue is how to define the importance of human generated rules. Different rules might have different impacts on task similarity measure and a methodology is required for handling this. Moreover, large scale evaluation is demanded to improve the proposed similarity algorithms.

## Acknowledgements

This work is supported by the MATURE IP funded by European Union Framework 7.

## References

1. Bergmann, R.: Experience Management. LNCS (LNAI), vol. 2432. Springer, Heidelberg (2002)
2. Sun, Z., Finnie, G.: Experience Management In Knowledge Management. In: Khosla, R., Howlett, R.J., Jain, L.C. (eds.) KES 2005. LNCS (LNAI), vol. 3683, pp. 979–986. Springer, Heidelberg (2005)
3. Wiig, K.: People-Focused Knowledge Management. Elsevier, Burlington (2004)

4. Riss, U.V., Grebner, O., Du, Y.: Task Journals as Means to Describe Temporal Task Aspect for Reuse in Task Patterns. In: Proceedings of the Ninth European Conference on Knowledge Management, UK, pp. 721–729 (2008)
5. Du, Y., Riss, U.V., Chen, L., Ong, E., Taylor, P., Patterson, D., Wang, H.: Work Experience Reuse in Pattern Based Task Management. In: Proceedings of the 9th International Conference on Knowledge Management, I-KNOW 2009, Graz, Austria, pp. 149–158 (2009)
6. NEPOMUK Deliverable D3.1.: Task Management Model (2007), <http://nepomuk.semanticdesktop.org/xwiki/bin/view/Main1/D3-1>
7. Tversky, A.: Feature of similarity. *Psychological Review* 84, 327–352 (1977)
8. Tversky, A., Gati, I.: Studies of Similarity. In: Rosch, E., Lloyd, B. (eds.) *Cognition and Categorization*, pp. 79–98. Lawrence Erlbaum Associates, Hillsdale (1978)
9. Jaccard Similarity, <http://www.dcs.shef.ac.uk/~sam/stringmetrics.html#jaccard>
10. Winograd, T.: Architectures for Context. *Human-Computer Interaction* 16, 401–419 (2001)
11. Huang, H., Gartner, G.: Using Activity Theory to Identify Relevant Context Parameters. In: Garther, G., Rehrl, K. (eds.) *Location Based Services and TeleCartography II. LNGC*, p. 35. Springer, Heidelberg (2009)
12. Basili, V.R., Caldiera, G., Dieter, R.: The Experience Factory. *Encyclopedia of Software Engineering* 2, 469–476 (1994)
13. Althoff, K.D., Decker, B., Hartkopf, S., Jedlitschka, A., Nick, M., Rech, J.: Experience Management: The Fraunhofer IESE Experience Factory. In: Proceedings of Industrial Data Mining Conference, Leipzig (2001)
14. Riss, U.V., Maus, H., Aalst, W.: Challenges for Business Process and Task Management. *Journal of Universal Knowledge Management* 0, 77–100 (2005)
15. Moody, P., Gruen, D., Muller, M.J., Tang, J., Moran, T.P.: Business activity patterns: A new model for collaborative business applications. *IBM System Journal* 45(4), 683–694 (2006)
16. Holz, H., Rostanin, O., Dengel, A., Suzuki, T., Maeda, K., Kanasaki, K.: Task-based process know-how reuse and proactive information delivery in TaskNavigator. In: Proceedings of the 15th ACM International Conference on Information and Knowledge Management, pp. 522–531. ACM Press, New York (2006)
17. BrainFiler home page, <http://brainbot.com/site3/produkte/brainfiler>
18. Shkundina, R., Schwarz, S.: A similarity measure for task contexts. In: Muñoz-Ávila, H., Ricci, F. (eds.) *ICCBR 2005. LNCS (LNAI)*, vol. 3620, Springer, Heidelberg (2005)
19. Sauer mann, L., Dengel, A., Elst, L., Lauer, A., Maus, H., Schwarz, S.: Personalization in the EPOS project. In: Proceedings of the Semantic Web Personalization Workshop at the ESWC Conference (2006)

# Behavioural Rule Discovery from Swarm Systems

David Stoops, Hui Wang, George Moore, and Yaxin Bi

University of Ulster, Jordanstown,  
Shore Road, Newtownabbey, County Antrim,  
Northern Ireland, BT37 0QB  
Stoops-d1@email.ulster.ac.uk,  
{h.wang,g.moore,y.bi}@ulster.ac.uk

**Abstract.** Rules determine the functionality of a given system, in either natural or man-made systems. Man-made systems, such as computer applications, use a set of known rules to control the behaviours applied in a strict manner. Biological or natural systems employ unknown rules, these being undiscovered rules which are more complex. These rules are unknown due to the inability to determine how they are applied, unless observed by a third party. The swarm is one of the largest naturally observed systems, with bird flocks and ant colonies being the most notable. It is a collection or group of individuals who use behaviours to complete a given goal or objective. It is the aim of this paper to present rule discovery methods for the mining of these unknown rules within a swarm system, employing a bird flock simulation environment to gather data.

**Keywords:** Rule Discovery, Behaviours, Swarm Systems, Data Mining, Artificial Intelligence.

## 1 Introduction

Operations carried out by many of the applications and systems within the real-world environment employ rules to determine how they act to given situations. The types of rules employed can be seen as either unknown, or known. A known rule is predefined to carry out an action upon state change; a user of the application will know what reaction will occur when using it. An unknown rule is not strictly defined; it can be seen within many biological or natural systems, with rules constantly being employed but with minimal understanding of how they react.

Known rules are employed by “man-made” systems such as those seen in computer applications; these have been written to carry out a given instruction when a state changes [1]. Biological and natural systems are “real-world” systems; these are generally seen as a set of rules employed by a biological entity such as people or insects. This type of system employs unknown rules, with this definition being used due to the properties of the behaviours seen not being fully understood. When considering the unknown rules groups within the real world which employ this rule type become prominent, namely the insect colonies and bird flocks. Each of these employ a set of rules which governs their behaviours, and each individual employs its own unique set

of rules based upon species and rank within its group [2-6]. This paper is concerned with the discovery of the individual rules which can be discovered from the collective group, or swarm.

Several issues arise within this particular problem, ranging from data collection and transformation, to the use of the classifiers to fully determine the most feasible solution for rule discovery. A simulation environment based upon the BOID Bird Flock application is to be used for data collection, with each individual being examined to determine the set of rules employed. Further to this, the group rules will be determined and individuals profiled within the system. It is envisaged that the rule discovery process be able to use a set of generic algorithms, allowing for the determination of rules within any system based upon a generic set of attributes. Application of this developed framework is to be used within any system with minimal alteration, and be able to determine the rule set employed. An additional concept used is to provide a means for profiling the individuals to further classify them by grouping, allowing for a more concise swarm discovery system.

Discovery of rules from swarms can further determine the use of their interactions through the behaviours employed. Although the discovery process is applied to the swarm as a whole, we are interested in each individual's behaviours. This leads to the further ability to determine the role of each member of the group, and further determine their group interactions.

This paper will present the developments within this area, the application of rule discovery methods within swarm data, and the results collected from each. This will include the presentation of the collected data set, the data transformation, and classification results.

## 2 Swarm Systems

Within the biological world there are groups of same species animals and insects, these are referred to as swarms [3-5]. Using these biological principles researchers have invested much into the understanding of how they work, and implementing these concepts into unique and novel application areas. Much work and time has been invested on this application area, with some Rule Discovery (RD) and data mining concepts being derived, namely the Ant Colony Optimiser and Fish School System [7-9]. The "*designing like nature*" approach within computing has meant that some of these biologically inspired systems have furthered understanding of their natural world counterparts [3].

A secondary application area has seen the true implementation of the behaviours applied to Artificial Life concepts. Similarly, the concepts of the swarm have been developed by NASA to create the ANTS (*Autonomous Nano Technology Swarm*) space exploration project [10]. This is based upon the Ant or Bee swarm/colony, which implements a hierarchy or rank system as seen in the biological counterpart. Within the natural world insect ecology there exists a "*chain of command*", with the Queen at the top, and the scouts or workers at the base level. This highlights the levels of interactions by the group, and therefore defines the set of rules employed by each member of the swarm according to their particular role.

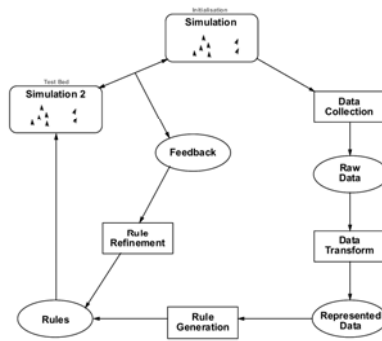
Furthering the Artificial Life concepts within this research field has also become quite prominent [5]. Early work saw the development of several simulated flocks (*groups, swarms and colonies*) based upon an observed rule set within a bird population. This is referred to as the BOID simulation, or Bird Flock simulation, which implements the observed set of rules, allowing for a simulated environment displaying the behaviours of the individual and group.

Swarm systems have attracted much interest; however the discovery of rules from swarms has seen little interest. The discovery of rules from swarms could provide insights into the behaviours and movements displayed within the natural world.

### 3 Discovery of Rules from Swarms

Discovery of rules draw from concepts used within Artificial Intelligence (AI). Approaches used can be defined as employing exploratory methods to discover useful and meaningful rules from data sets. The aim of employing the AI methods and algorithms is to allow for data sources to be mined, with a strong set of rules being discovered.

Previous research into this field led to a framework being developed, providing a generic method for the discovery of rules from a predefined set of attributes. Fig 1. highlights the segment of the framework which is to be discussed within this paper, presenting the process required to mine for the rules.



**Fig. 1.** Rule Discovery generic framework for the processes required to mine for the BOID simulation rules

As previously discussed, little work has been carried out within the field of rule discovery from swarms. A previous study carried out used the M5 Rule classifier, a regression tree classification algorithm which employs a separate-and-conquer approach for the discovery of rules. This paper aims to present alternative methods for the discovery of rules from the data set, using a further refined set of attributes which define the requirements for the rule set required.

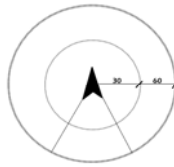
Fig 1. highlights the implementation of the simulation, with data collection, data transformation, rule discovery mining, and, re-implementation of the discovered rule

set. Within the following sections this paper will present the refined attribute selection, findings of the study, and the rules discovered during the process.

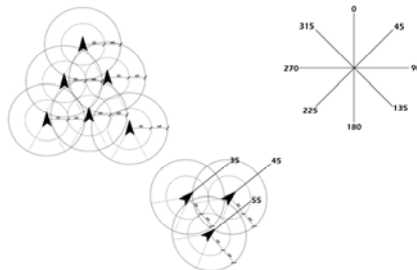
### 3.1 Simulation Environment

The BOID, or Bird Flock simulation environment is used as the base for the data collection on swarm behaviours and interactions. The behaviours and interactions will define the rules employed within the system. The main rules are the behavioural rules which are defined as cause-and-effect, leading to actions being triggered upon a change of state. The interactions also define some parts of the discovered rules with the individuals interacting with other members of the swarm, triggering actions to be taken.

The simulation is comprised of a set boundary space for the environment, employing a set of represented birds (*see Fig 2. for representation*), each coloured by species or grouping, and a set of rules employed by each individual. The assumption made for this study is that all species (*by colour*) employ the same set of rules to guide their behaviours. Several constants are present for each species of the BOID/Bird, this being their visual range, detection distance (*proximity to another BOID/Bird to detect presence*), and a neighbourhood distance (*in this case, distance is wingspan with no other individual occupying the same space*). Fig 2. highlights these constants, with the range values being different per species.



**Fig. 2.** Visual representation of the BOID/Bird, and the constants for each individual



**Fig. 3.** Swarm/Flock representing the neighbourhood ranges for the individual and swarm, and the Theta directional heading for each instance

A similar constant employed within the simulation for the BOID representation is the Theta value; this determines the angle by direction in which the BOID is travelling. Theta only determines the directional angle for each instance, therefore, if there are 3 instances in a group, each will have a different directional heading. Fig 3. presents this concept, highlighting the swarm neighbourhood and the Theta headings.

The rules presented within the simulation are those which were defined by Reynolds research into the BOID simulation, and the assumptions made when considering the rules to define [4]. The simulation will employ this rule set, and from the running version data will be collected. The 3 rules determined by Reynolds are [4]:

**Alignment:** defined as the rule which governs the directional heading of all individuals within the swarm, ensuring that all individuals are maintaining the same heading.

**Cohesion:** defined as the rule that governs the proximity of the individuals within the swarm, ensuring that all individuals who are within the detection range of another member maintains their distance without becoming separated.

**Separation:** defined as the rule which ensures that all individuals maintain a “safe” distance from their neighbouring member. The individuals must not enter the neighbourhood range, if they do, they will change heading until outside of the range then match group heading.

When considering these rules as a pseudo-code representation the rules can be written and presented as:

**Alignment:** IF nearest\_theta\_change > 0  
                   THEN theta\_change = nearest\_theta\_change;  
                   speed = nearest\_speed

**Cohesion:** IF nearest < 60  
                   AND nearest > 30  
                   THEN theta = nearest\_theta;  
                   speed = nearest\_speed

**Separation:** IF nearest < 30  
                   THEN nearest = nearest + (nearest \* 0.1);  
                   theta = theta + 1;  
                   speed = nearest\_speed

The current study will use these 3 rules as a basis for comparison with the discovered rules. In the following sections we will look at the data collection, transformation, and rule discovery.



## 4 Data Generation

The aim of this work is to provide a generic method for discovering rules from swarm systems. Research into the features and attributes required to effectively and correctly discover rules have provided a set which can provide the best outcomes. The data which is to be collected from the simulation environment can be defined as the minimum required attributes that can be used to discover the required rules.

The data collection process defined in the previous study employing the M5 Rule classifier has been further refined to include the most feasible attributes [1][11]. Table 2. presents sample data collected from the simulation in its raw format, this will be transformed into a minable format.

**Table 1.** Sample data collected from the BOID/Bird Flock simulation, covering 3 instances and 3 species

System Time	Instance ID	X Coord	Y Coord	Theta	Speed	Neighbour	Detect
1.27E+12	0	544	196	128	5	30	60
1.27E+12	1	974	517	3	5	30	60
1.27E+12	2	720	598	72	5	30	60
1.27E+12	3	973	491	359	5	40	80
1.27E+12	4	964	554	10	5	40	80
1.27E+12	5	735	561	73	5	40	80
1.27E+12	6	0	481	353	5	50	100
1.27E+12	7	944	299	170	5	50	100
1.27E+12	8	88	125	300	5	50	100

The data presented in Table 1. shows the raw format for the instances within the simulation. The use of 3 species provides alternative data from which to derive the rules for each of the independent groups, and rules for each of the individuals within each of the groupings. Without gathering any additional data on each of the BOID's (*for example, colour*) the aim is to discover the rules, with the set of rules being discovered to profile each individual and their species (*colour*).

Data transformation steps are undertaken in order to provide the more minable data set. The whole data set is processed with all individuals recorded, and transformed as a group. The data mining process requires that the rules are discovered from the group over a given time period, in this case, the time period undertaken is a 10 minute segment (*36,000,000 milliseconds*). Throughout this given time the script records a reading of the attributes defined in Table 1. every 200 milliseconds. It has been

determined, that for an accurate reading of the group activities and behavioural activation, this time period is valid. A behaviour can be activated within a short period of time, and as such must be recorded as, and when it is activated.

The distance measurements (*pixels*) shown in the tables have been derived from using the two positions of the BOID individual. Application of the Euclidean distance measurement has been employed to determine the straight line distance between the two points in time using the XY co-ordinates. The Euclidean distance measure can be used within the two-dimensional or three-dimensional spaces, further providing methods for the incorporation of a generic rule discovery process.

From simulation activation, the data is captured as discussed, and transformed as shown within Table 2.

**Table 2.** Sample of transformed data from the BOID/Bird Flock simulation as required for Rule Discovery mining

ID	Class	Time (ms)	Theta	Theta Change	Speed	Near Speed	Near	Neighbour	Detect
0	BOID	200	128	1	5	5	200	30	60
1	BOID	200	3	0	5	5	81	30	60
2	BOID	200	72	0	5	5	106	30	60
3	BOID	200	359	0	5	5	201	40	80
4	BOID	200	10	0	5	5	98	40	80
5	BOID	200	73	0	5	5	302	40	80
6	BOID	200	353	0	5	5	105	50	100
7	BOID	200	170	0	5	5	208	50	100
8	BOID	200	300	0	5	5	68	50	100

The attributes selected for mining refer to the rules already defined within the original simulation. It has been determined that the basis for these rules can be successfully represented within other simulations or environments for rule discovery mining. These attributes can be applied within the current two-dimensional environment, and also to a three-dimensional environment. This is due to the detection range value which defines the proximity to another individual. If a third dimension is added it will not be detected if the other members of the groups are on other floors or on a higher/lower plane.

Important attributes for the discovery process refer to the directional heading, change in direction, speed (*individual and nearest neighbour*), and the nearest neighbour instance. Generally the rules which will be discovered relate to the movements of the individuals within the group, additional rules which are not perceived may be derived based upon these movements. With the swarm we are interested in

discovering the interactions between individuals upon joining a group, or when already part of a group.

To determine which state is currently being employed by the individual we determine its proximity to other instances, the nearest neighbour providing the relevant data. This then gathers its movement information, including the speed it is currently moving at (*pixels/second*), and the speed which its nearest neighbour is moving at. Similar importance is given to the heading of the individual; however, this is only relevant to the current individual. This is due to the differences in theta captured by the system, each individual within the swarm will have a different theta heading. This leads to the detection of the change in theta by the individual, if a change occurs this can be detected throughout all members of the group.

Table 3. highlights the data for an individual within the swarm, the data has been processed as previously discussed.

**Table 3.** Sample of transformed data for an individual within the group, for a single species

ID	Class	Time (ms)	Theta	Theta Change	Speed	Near Speed	Near	Neighbour	Detect
0	BOID	1000	99	1	5	5	34	30	60
0	BOID	1200	98	1	5	5	34	30	60
0	BOID	1400	97	0	5	5	34	30	60
0	BOID	1600	97	0	5	5	35	30	60
0	BOID	1800	97	0	5	5	35	30	60
0	BOID	2000	97	0	5	5	35	30	60
0	BOID	2200	97	0	5	5	34	30	60
0	BOID	2400	96	1	5	5	34	30	60
0	BOID	2600	96	1	5	5	34	30	60

The data in Table 3, details the movement for a single instance. The nearest instance remains within the 30-pixel boundary area between the neighbourhood and detection distance, and generally closer to the neighbourhood range. The changes in theta can be seen as a single pixel at a time, this can be seen as minor adjustments by the individual within a group. A higher value would be a clear indication that the individual has joined a larger group, leading to the larger change in direction.

Similarly, when reviewing the transformed data within the other species data sets we can see a similar trend. The nearest neighbour tends to remain closer to the neighbourhood range, rather than the detection range. This can be seen with the other values within the data, small changes in the theta are seen as the adjustments of the individuals within the group, and large ones can be viewed as joining a larger swarm grouping.

The application of the M5 Rule classifier aims to determine the rules employed by the swarm as a whole, and define an overall rule-set employed. Further to this we aim to use these discovered rules to profile each individual within the swarm, allowing for individuals to be classified further into a specific role. This can be considered when looking at swarms such as bee or ant colonies, where there are multiple rules with each employing their individual rules.

Within the following sections we will present the results gathered from the rule discovery process, and discuss the relevance of these to the work in general.

## 5 Discovered Rules

A previous study involved the use of the M5 Rule classifier, employing the regression tree model to discover rules from the data set [1][11][12]. The M5 Rule regression tree classifier uses the separate-and-conquer approach at each best leaf node to determine the strongest possible output rules [11]. Changes to the data transformed have led to a need to reinvestigate the set of rules discovered by this classification method. The previous study provided a set of rules relevant to the BOID simulation, however initial work with other types of swarm simulation, in this case a simplified bee swarm, did not provide the strength of rule envisaged for this generic discovery process. An evaluation of the previous study combined with the outcomes from the bee swarm provided differing results, this led to a deeper review of the attributes employed within both swarms, leading to the process being refined to define a stricter set of attributes for the discovery process. It is foreseen that a generic framework application can be developed and applied to further swarm systems. The same set of attributes is required throughout the discovery process for all types of swarm, with the strength of rules being maintained. Further work will employ measures for rule strength, enhancing the discovery process.

Further to this investigation the strength of rule discovered has improved, providing only the 3 rules which are implemented. An additional rule has been discovered during this process, relating to the BOID's natural state. Below are the rules discovered using the M5 classifier and extended attribute set.

**Alignment:** IF nearest\_theta\_change > 1  
 THEN theta\_change = theta\_change + 1;  
 speed = 5;  
 nearest >= 30.85;  
 nearest <= 60

**Cohesion:** IF nearest < 60  
 AND nearest > 30.85  
 THEN theta = theta + 1;  
 speed = 5

**Separation:** IF nearest < 30.85  
 THEN nearest = nearest \* 0.12;  
 theta = theta + 5;  
 speed = 5

The additional rules which was discovered using the M5 classifier declared that any nearest > 60, led to speed and theta remaining the same.

**Typical state:** IF nearest > 60  
 THEN speed = speed;  
 theta = theta

The M5 classifier provided strong results, the improvement and addition to the set of required attributes has led to greater accuracy within the discovery process. Building on this example we now present the use of the M5P Tree, or M5 pruned tree [11][12]. Below are the results gathered from the M5P Tree are presented below.

**Alignment:** nearest > 60  
 nearest < 30  
 THEN  
 speed = speed \* 1;  
 theta\_change = theta\_change + 1;  
 theta = theta + (theta \* 0.5);

**Cohesion:** nearest > 60  
 nearest < 30  
 THEN  
 speed = speed \* 1;  
 theta = theta \* 1;

**Separation:** nearest < 30  
 THEN  
 Theta = theta + 1;  
 Speed = 5;  
 Nearest\_dist = nearest \* 1.5

The discovered rules are provided for a single species within the swarm simulation, the simulation was carried out using 3 species employing differing attribute values. The highlighted rules a centred around a single species where the Speed is a constant value. The ‘Theta’ and ‘Theta\_Change’ values use the directional heading. ‘Theta\_Change’ is ‘0’ when the heading for the BOID is maintained as a straight path heading, and is greater than ‘1’ when it employs obstacle deviation or joins a larger group. There are only a few occurrences of the ‘Theta\_Change’ value being greater than ‘1’, with this larger state change only occurring once for each BOID as it joins a larger grouping.

## 6 Discussion

Upon reviewing the results collected from the rule discovery mining process, there are several observations. With the work carried out when reinvestigating the attributes used to mine from, we can see that the accuracy of the discovery process has

improved. The final step of the process is to develop the rules into a code fragment and apply it to the simulation to determine if it is functional data. Further work will see the implementation of a more refined refinement process, involving measurements of accuracy.

The process employed for the discovery of rules from swarms is designed as a generic application for discovery from other types of swarm systems. The generic framework has been applied to the BOID Bird flock simulation, and initially applied to a simple bee swarm; this led to very differing results. The bee swarm result provided many weak rules, whereas the BOID simulation provided stronger results. This required the initial attribute selection to be revisited, further defining the important attributes and their application to each type of swarm.

Two important factors which impact upon the discovery process involve the data itself. Firstly the number of records collected for each instance during the 10 minute time period is approximately 7,200,000 records. This impacts greatly upon the accuracy of the discovery process, leading to the need for a further refinement process. The second factor again relates to the data collected. The time required to process and transform the number records can be substantial, with a large number of records per instance, it increases processing time greatly per instance.

The complexity of the data and number of records is not a major factor for the M5 Rule classifier; however it does affect the accuracy and classification within the M5P Tree. The higher the number of records being processed, the less accurate the M5P becomes, leading to a less accurate method for classifying and discovery trends and rules. The requirement is to incorporate a robust method for rule discovery, and to provide the most accurate method available. The importance is on providing an algorithm which can be easily built into a generic framework system.

The M5 Rule classifier provides the most appropriate solution for this style of framework, it provides an accurate set of rules, that when applied back into the simulation behaved in a similar manner to the original version. This classifier also correctly discovered the ranged values of each species within the data set, which the M5P Tree was unable to fully achieve. This is an important aspect within this particular area of research, each species maintains its own set of rules, even though we assume that for this simulation, they remain the same rules with minor alterations. Within a profiling system we require the rules to be discovered based upon the species, allowing for them to be profiled by species as well as role.

Initial work developing a profiling system has been built to differentiate between the species when classifying their roles. Roles are defined by the rules which are used most within their data collected, providing a set of rules employed and therefore setting their position in the group. Complexity of the swarm hampers the occurrence of some rules, meaning that not all rules are used by an individual unless they are activated, this drawback can mean that some behaviours are not discovered.

These factors may lead to the requirement of running the simulation several times to attempt to trigger as many of the states as possible. This should lead to the full rule set being discovered.

## 7 Further Work

Future work in this area will see the development of the Rule Discovery process, creating a more robust version. This will see the inclusion of other types of algorithm in order to determine the most effective. A study using a Genetic Algorithm is planned, using this method to further determine a range of algorithms and their particular application to the discovery of rules from swarms. Work has been started with the implementation of the profiling element of the system, leading to a further refinement of the profiling using the discovered rules and data collected. This will provide further value to the discovery process, leading to a process of discovering rules and determining the roles of each individual within the swarm, or swarms if there are multiple species or groupings.

## References

1. Stoops, D., Wang, H., Moore, G., Bi, Y.: Rule Discovery from Swarm Systems. In: Proceedings of the Eighth International Conference on Machine Learning and Cybernetics, pp. 3544–3549 (2009)
2. Flake, G.W.: *The computational beauty of nature*, 1st edn. MIT Press, Cambridge (1998)
3. Kennedy, J., Eberhart, R.: *Swarm Intelligence*, 1st edn. Morgan Kaufmann, San Francisco (2001)
4. Reynolds, C.W.: Flocks, Herds, and Schools: A Distributed Behavioral Model. *Computer Graphics* 21(4), 25–34 (1987)
5. Adami, C.: *Introduction to Artificial Life*, 1st edn. Springer, Berlin (1998)
6. Dorigo, M.: Swarms. *Swarm Intelligence* 1(1), 1–2 (2007)
7. de Almeida Prado, G., Toracio, A., Pozo, A.T.R.: Multiple objective particle swarm for-classification-rule discovery (2007)
8. Zhang, M., Shao, C., Li, M., Sun, J.: Mining Classification Rule with ArtificialFish Swarm (2006)
9. Bo, L., Abbas, H.A., McKay, B.: Classification rule discovery with ant colony optimization (2003)
10. NASA, Autonomous Nano Technology Swarm, NASA ANTS Homepage, <http://ants.gsfc.nasa.gov/>
11. Hall, M., Holmes, G., Frank, E.: Generating Rule Sets from Model Trees. In: Foo, N.Y. (ed.) *AI 1999. LNCS*, vol. 1747, pp. 1–12. Springer, Heidelberg (1999)
12. Witten, I.H.: *Data Mining: practical machine learning tools and techniques*, 2nd edn. Morgan Kaufmann, London (2005)

# Knowledge Discovery Using Bayesian Network Framework for Intelligent Telecommunication Network Management

Abul Bashar<sup>1,\*</sup>, Gerard Parr<sup>1</sup>, Sally McClean<sup>1</sup>,  
Bryan Scotney<sup>1</sup>, and Detlef Nauck<sup>2</sup>

<sup>1</sup> School of Computing and Engineering, University of Ulster,  
Coleraine BT52 1SA, UK

{bashar-a, gp.parr, si.mcclean, bw.scotney}@ulster.ac.uk

<sup>2</sup> Research and Technology, British Telecom, Adastral Park,  
Ipswich IP5 3RE, UK  
detlef.nauck@bt.com

**Abstract.** The ever-evolving nature of telecommunication networks has put enormous pressure on contemporary Network Management Systems (NMSs) to come up with improved functionalities for efficient monitoring, control and management. In such a context, the rapid deployments of Next Generation Networks (NGN) and their management requires intelligent, autonomic and resilient mechanisms to guarantee Quality of Service (QoS) to the end users and at the same time to maximize revenue for the service/network providers. We present a framework for evaluating a Bayesian Networks (BN) based Decision Support System (DSS) for assisting and improving the performance of a Simple Network Management Protocol (SNMP) based NMS. More specifically, we describe our methodology through a case study which implements the function of Call Admission Control (CAC) in a multi-class video conferencing service scenario. Simulation results are presented for a proof of concept, followed by a critical analysis of our proposed approach and its application.

**Keywords:** Next Generation Networks (NGN), Network Management, Bayesian Networks (BN), Call Admission Control (CAC).

## 1 Introduction

The area of telecommunication networks is changing at a rapid pace in terms of its architecture and services due to the advances in the underlying technology and also owing to the new demands placed by the consumers who use them. The latest dominant technology in this domain is the Next Generation Network (NGN), which is capable of providing converged services with guaranteed QoS whilst offering enormous savings to the network and service providers by reducing their Capital Expenditure (CAPEX) and Operational Expenditure (OPEX) [1]. The majority of the existing Internet traffic is carried over a combination of wireline and wireless communication infrastructure. This infrastructure includes customer premises, telephone exchanges, base stations and core trunk network links.

---

\* Corresponding author.



In essence, the core network of NGN involves a consolidation of several transport networks, each built for a different service, into one core transport network based on IP (Internet Protocol) and MPLS (Multi Protocol Label Switching). The access network is a highly complex mix of wired and wireless technologies, including ADSL (Asymmetric Digital Subscriber Line), FTTP (Fibre-To-The Premises), WLAN (Wireless Local Area Network), WiMAX (Worldwide Interoperability for Microwave Access), etc. The edge network connects the multi-user access network to the high speed core network to complete the NGN architecture. From the architectural, functional and management points of view, NGN has become a highly complex, dynamic and unpredictable system which the traditional Network Management Systems (NMSs) find difficult to control and manage [2]. This situation has led researchers to explore intelligent and autonomic approaches for improving the functionality of NMSs so that they can operate in a more efficient manner to achieve the desired management goals and objectives.

Machine learning (ML) is one such approach which has the capability to address the issues of improving the network management functionalities by imparting automated and intelligent data analysis techniques to network management datasets. These techniques have the ability to learn the system behaviour from past data and estimate future behaviour based on the learned system model [3]. In this paper we present the application of a graphical modelling technique, Bayesian Networks (BN), to assist the NMS in solving performance management problems in a network. The original contribution of this paper is the application of BN to model the behaviour of network elements (e.g., the routers) and to provide a mechanism to perform probabilistic inference, decision making and prediction to achieve desired QoS goals through the process of Call Admission Control.

The remainder of this paper is structured as follows. In Section 2 we provide the required background and survey some related work in this research domain. Section 3 describes the theory behind BN and the learning algorithms utilised in our solution. In Section 4 we present the conceptual framework of our generic approach using BN-based DSS. We then demonstrate the proof of concept through simulations using *OPNET* and *Hugin* tools in Section 5. The results of the simulation, BN model validation and related discussions are presented in Section 6. Section 7 concludes the paper by suggesting possible future work.

## 2 Related Work

The task of a NMS is to monitor and control the functionalities and behaviour of various constituent network elements so as to achieve a desired system state, given a set of pre-defined high-level business goals. The goal could be to maximise resource utilisation (translated as maximising revenue to service providers) or minimise service response times (equivalent to enhanced customer satisfaction). To achieve these objectives, there is a requirement for a standardised and efficient NMS. Several standard NMSs like the SNMP [4], CMIP and FCAPS (Fault, Configuration, Accounting, Performance and Security) [5] are used to manage the telecommunication networks. However, since NGN is a converged

network providing data and voice services, it would have to be managed by the combination of existing NMSs. One aspect which is common to all NMSs is that they deal with collections of management data typically stored in Management Information Bases (MIBs) at network nodes. To analyse and interpret such massive data sets by a human operator would be almost an impossible task. Hence, we need ML techniques to automate such tasks by utilising rich reasoning mechanisms to find patterns which are useful for network management.

Various ML approaches have been proposed to address the issues related to network management functions. Decision trees have been used to achieve proactive network management by mining the data obtained from SNMP MIB objects [6]. Bandwidth broker design can be facilitated by applying Fuzzy Logic [7]. Predictive network fault detection in a telecommunication network has been addressed using Bayesian Belief Networks (BBN) [8]. Dynamic Bayesian Networks (DBN) have also been used for detecting network faults in real-time [9]. Bayesian reasoning based software agents have been used to achieve intelligent network fault management [10]. Reinforcement learning has been used to provide efficient bandwidth allocation in Differentiated Services (Diffserv) networks for per hop behaviour aggregates [11]. Router performance modelling has been achieved using the learning features of Bayesian Belief Networks [12] [13]. An interesting survey on applications of AI techniques to the telecommunications domain that summarises over a decade of research work, is provided in [14].

Based on this survey, it is seen that the majority of research has concentrated on applying ML to the fault management function of the broader network management domain. Also, it is seen that BN are suitable for modelling and studying highly dynamic systems. We now consider the application of BN to implement and enhance the performance management function, which is one of the key functions of management in the FCAPS model of ITU-T [5]. The performance management function which we will be addressing is Call Admission Control and it will be described later in Section 5.

### 3 Bayesian Networks Theory

#### 3.1 Graphical Structure of Bayesian Networks

A BN is a graphical structure that allows us to represent and reason about an uncertain domain. For a set of variables  $X = \{X_1, \dots, X_n\}$ , a BN consists of a network structure  $S$  that encodes a set of conditional independence assertions about variables in  $X$ , and a set  $P$  of local probability distributions associated with each variable [15]. An edge from one node to another implies a direct dependency between them, with a child and parent relationship. To quantify the strength of relationships among the random variables, a conditional probability function  $P$  is associated with each node, such that  $P = \{p(X_1|\Pi_1), \dots, p(X_n|\Pi_n)\}$ , where  $\Pi_i$  is the parent set of  $X_i$  in  $X$ . If there is a link from  $X_i$  to  $X_j$ , then  $X_i$  is a parent of  $X_j$  and thus it belongs to  $\Pi_j$ . For discrete random variables the conditional probability functions are represented as tables, called Conditional Probability Tables (CPTs). For a typical node  $A$ , with parents  $B_1, B_2, \dots, B_n$ ,

there is associated a CPT given by  $P(A|B_1, B_2, \dots, B_n)$ . For root nodes, the CPT reduces to prior probabilities. The main principle on which BN work, is Bayes' rule:

$$P(H|e) = \frac{P(e|H)P(H)}{P(e)} \quad (1)$$

where  $P(H)$  is the prior belief about a hypothesis,  $P(e|H)$  is the likelihood that evidence  $e$  results given  $H$ , and  $P(H|e)$  is the posterior belief in the light of evidence  $e$ . This implies that belief concerning a given hypothesis is updated on observing some evidence.

### 3.2 Features of Bayesian Networks

*Structural Learning:* The structure of the BN can be constructed manually by the subject expert or through structure learning algorithms - PC (Path Condition) and NPC (Necessary Path Condition) algorithms [16] [17]. The basic idea of these constraint-based algorithms is to derive a set of conditional independence and dependence statements (CIDs) by statistical tests among the nodes of the BN.

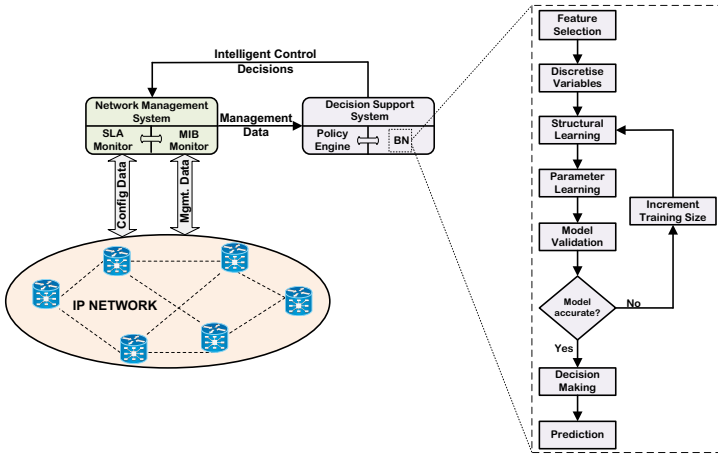
*Parameter Learning:* The CPTs (or parameters) can be specified, based on the knowledge of the domain expert, by the process of parameter elicitation. The past data may also be used as the basis for learning the parameters using efficient algorithms. The Expectation Maximization (EM) algorithm is particularly suitable for batch parametric learning [18], while Adaptation algorithms are useful for sequential parameter updates [19].

*Inferencing:* Evidence on a particular node is used to update the beliefs (posterior probabilities) of other nodes of the BN. The BN framework supports predictive and diagnostic reasoning and uses efficient algorithms for this purpose [20].

*Decision-making:* To incorporate decision making capabilities, the BN is converted to an influence diagram (ID) by adding decision nodes and utility nodes. The values taken by the decision nodes inform the actions which must be chosen by the decision maker. A utility node quantifies the *usefulness* of the outcomes resulting from the actions of decision.

## 4 A BN-Based Decision Support System

The domain under consideration for our proposed work is shown in Fig. 1. It consists of two modules, namely, the Network Management System (NMS) and the Decision Support System (DSS). The NMS is assumed to be based on the SNMP protocol, which collects network management data using the SNMP Management Information Base (MIBs) of the network elements. MIB is a virtual database for the managed entities which is defined at various layers of protocol stack and provides management information. In our case, we wish to monitor two data sets, namely MIB data pertaining to the IP layer (Incoming packets,



**Fig. 1.** Conceptual framework for the BN-based DSS

Outgoing packets) and SLA (Service Level Agreement) data to check the service level agreements pertaining to the QoS metrics such as Delay, Packet Loss and Jitter. The collected data are then fed from the NMS to the DSS, where the latter builds a model of the network behaviour using the BN framework.

The details of the steps involved in the BN model construction are also shown in Fig. 1 through a flowchart. First step is to identify the variables of interest through the process of feature selection. This step eliminates the variables which do not contribute to the model and hence reduces its complexity. Then the continuous variables are discretised into pre-defined number of states. The choice of the number of states is a compromise between modelling accuracy and system complexity. The learning phase involves structural and parameter learning as discussed in Section 3.2. To verify that the model is suitable for estimation, we perform model validation through the process of k-fold cross validation. If the model performs below a pre-defined performance threshold, we increment the size of training data set. Once the desired accuracy is reached, the model is used to make decisions and also to predict future behaviour. These decisions from the BN are then fed back to the NMS to achieve the desired high-level goals as pre-defined through the policy engine. Usually the network manager or the domain expert can make appropriate policies for the required objectives. After the decisions are fed back to the NMS, they are translated into configuration changes which need to be made to the physical network elements.

## 5 Case Study and Simulation Setup

### 5.1 Motivation

One of the main promises of NGN is the provisioning of guaranteed QoS. It is well known that even extremely well designed networks can suffer from performance

degradation (reduced QoS) due to the congestion problem. Call Admission Control (CAC) is a preventative approach to deal with congestion, which makes decisions to accept a new call, based on whether this new call can be supported with the desired QoS [21]. A detailed study of the current CAC techniques reveals that there are two major classes of CAC algorithms: Traffic-model-based and Measurement-based [25]. Traffic-model based schemes have the disadvantage that they do not consider the long-range dependence property (i.e., slow decay of the autocorrelation function) which is an important characteristic of the NGN traffic. Measurement-based approaches make admission decisions based on the current network status, which they obtain through periodic measurements of QoS metrics (e.g., end-to-end delay). Hence they can achieve much higher network utilisation and also provide desired QoS. We base our case study on the latter approach and improve it using the BN modelling technique. In our approach, we reduce the overhead of periodic measurements by using the QoS estimates generated from the BN model.

### 5.2 Simulation Setup

To demonstrate the proof of concept, we used commercial simulation software packages *OPNET* [22] (for telecommunication network simulation) and *Hugin Researcher* [23] (for BN modelling). Exchange of simulated data between the software systems was done using a C++ programming framework. The OPNET screenshot of the network topology used for the case study is shown in Fig. 2.

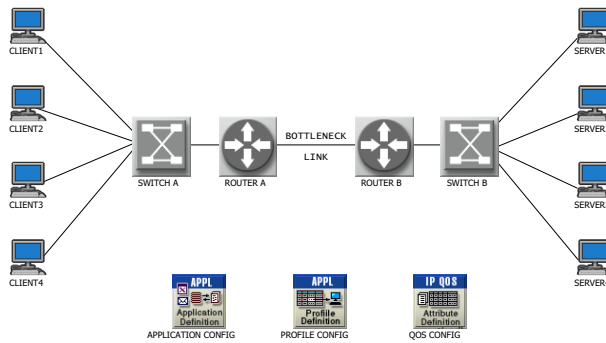


Fig. 2. Network topology for case study

We consider four clients and four servers connected through switches and routers. The application running on the network is video conferencing under varying TOS (Type Of Service). The router on which we focus is ROUTER A, (which connects to ROUTER B by the bottleneck link) where we capture the MIB data related to network statistics. At this router we implement the BN modelling and then make decisions based on the obtained model. The clients have been modelled with traffic characteristics as detailed in Table 1.

**Table 1.** Characteristics of Traffic Sources for Fig. 2

Source	Type of Service	Priority
Client1	Background	32
Client2	Standard	64
Client3	Excellent Effort	96
Client4	Streaming	128

The clients with particular TOS connect to a corresponding server operating under similar TOS. The TOS is determined based on the DSCP (Differentiated Services Code Point) [26] values and it affects the treatment of the packet in the router queues. From Table 1 it can be seen that each TOS has an assigned priority number in the range (0-252) with 0 being the lowest priority (best effort service) and 252 being the highest priority. The higher priority traffic gets through the router first in cases of congestion. The sources generated flows with exponential distributions to simulate traffic as observed in the NGN environment.

### 5.3 Experimental Details

We ran simulations and collected router MIB statistics at every 60s to obtain historic data. The choice of this sampling rate is inspired by the general practice of network managers who use this practical SNMP polling rate [24]. The details of the statistics collected are given in Table 2, where each statistic corresponds to a node in the BN.

**Table 2.** Network statistics which form the BN nodes (at ROUTER A)

BN Node	Description
Received	Traffic received at the router from traffic sources (bytes/sec)
Sent	Traffic sent from the router to next hop (bytes/sec)
Throughput	Successful packet transmission rate on the bottleneck link (bits/s)
Delay	Instantaneous value of packet waiting times in the queue (s)
Jitter	Packet delay variation in the queues (rate of change of delay)

One of the major operations was to discretise the collected data. The choice of discretisation levels determines the accuracy required for representing the collected data. It is to be understood that, the greater the number of levels, the larger the size of the CPTs, and hence a proportional increase in the model complexity. We chose five discretisation levels with equal bin sizes, and the levels were defined as Very Low (*VLO*), Low (*LO*), Medium (*MED*), High (*HI*) and Very High (*VHI*). After this process the data were fed into the Hugin Researcher for building the BN models using the algorithms discussed in Section 3.2.

Even though in a real network there are many MIB variables (in hundreds) which are monitored and collected, not all of them are relevant when building

the model. The importance of variables for the model can be determined by the feature selection process which allows us to make a good choice of variables for the prediction task. In our case we began with eight variables and found that only five of them were required for the model (the dropped variables were *Bit error rate*, *Bit errors per packet* and *Packet loss ratio*).

## 6 Simulation Results

We now present the results of the simulation for our case study. The results include the BN models which depict the structure and the parameters, sample admission control decisions and delay predictions made.

### 6.1 Effect of Number of Training Cases

Our first result shows (Fig. 3) the effect of the number of cases on the construction of the BN model. To check the accuracy of the model, we performed a 10-fold cross validation. We observed the prediction accuracy of all the BN nodes, but present here only the results for the *Delay* node. This is because we would like to make admission decisions based on the estimates of delay QoS. For this purpose we had to vary the simulation time and capture statistics (number of cases) at a fixed interval of 60s. We started with 100 cases and went up to 3000 cases. The criterion to achieve a suitable model was to get high prediction accuracy with low prediction error deviation. Fig. 3 presents the results of this process and we found that 2000 cases were sufficient to get a stable and accurate model, as increasing cases from 2000 to 2500 improves the prediction accuracy by less than 0.2% at the cost of 500 minutes of simulation time. Hence, we present the BN model for Router A (see Fig. 4), which was obtained from a training data size of 2000 cases.

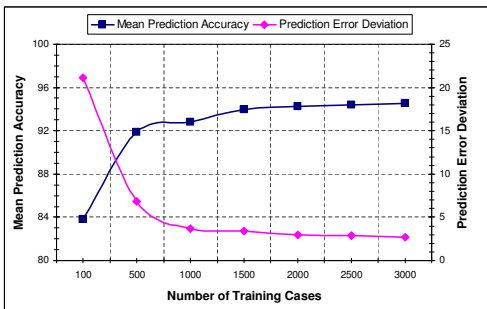


Fig. 3. Prediction accuracy plot

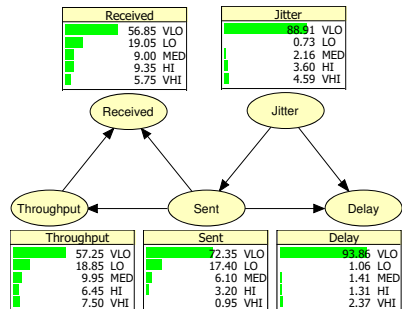


Fig. 4. BN for Router A

### 6.2 Decision Making Using Influence Diagrams

The second result shows how the BN can be used for making decisions under uncertainty. For this purpose, first we convert our BN model (see Fig. 4) into an Influence Diagram. Fig. 5 shows this transformation, where we can observe two extra nodes that are added. The diamond shaped node is called the utility node and the rectangular shaped node is called a decision node (with two actions: *Admit* or *Deny* traffic). The utility function defines the measure of goodness of a specific decision and is specified as a table (see Table 3) which quantifies the expert knowledge of making *correct* decisions. When combined with the observed evidence, this can help us to quantify our decisions in a particular situation. For example, in Fig. 5 the ID makes a decision to *Admit* a call because the reward of this action (59.4) is greater than the other action (*Deny* is 40.6) based on observation of a single piece of evidence (*Received* traffic is in *VHI* state). This decision can be compared with the decision made in Fig. 6, where we have more information in terms of two pieces of evidence (with additional evidence of *Jitter* value being in *VHI* state). In this situation the ID decides to *Deny* traffic admission, since now the reward for *Deny* action (80.1) is greater than the *Admit* action (19.9). This is intuitively a correct decision, because if jitter is high and at the same time there is high incoming traffic, then this traffic needs to be denied so as to respect the QoS constraints. In this way the network manager may take the support of the BN models to cross check the actions which need to be taken in a particular situation. As a matter of fact, we have checked the decisions made by our ID in various scenarios and verified that it does make correct decisions for call admission. However, due to space limitations we are unable to present them here.

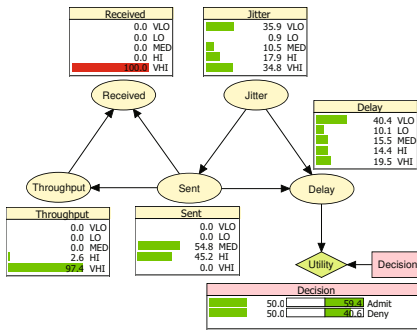


Fig. 5. Decision with one evidence

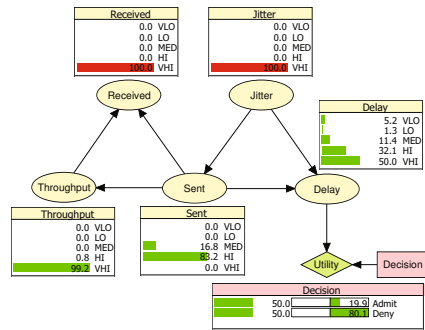


Fig. 6. Decision with two evidences

Table 3. Utility Table

Decision	Admit					Deny				
	Delay VLO	LO	MED	HI	VHI	Delay VLO	LO	MED	HI	VHI
Utility	100	75	50	25	0	0	25	50	75	100



### 6.3 Future Predictions of Delay

The final result of our case study is presented in Fig. 7, which shows the prediction capabilities of our BN model with respect to the *Delay* node variable. We tested four BN models (obtained with training cases of 500, 1000, 1500 and 2000) to predict *Delay* variable for two different future intervals of 500 cases (from time 2000s to 2500s) and 1000 cases (from time 2000s to 3000s). We assume here that *Delay* node is unobservable or prohibitively costly to observe. The first observation is that the prediction was better for 500 cases in future as compared to 1000 cases. This is because as more time elapses, the models tend to become *obsolete* and need to be updated. The second observation is that prediction accuracy improves with the model training size, which is to be expected. The maximum prediction accuracy was 96.85%, which is considered to be significant in terms of saving on the overhead of delay measurements (and instead using estimates from the BN model) for making admission control decisions.

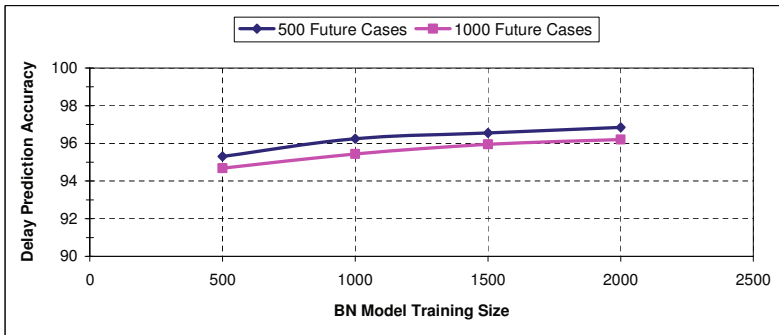


Fig. 7. Predicting *Delay* in the future

### 6.4 Discussion of Results

It has been demonstrated that BN can be used as inference, decision making and prediction technique for network management. We now critically evaluate our approach by considering the following issues.

1. *Practicality*: The procedure presented is practically realisable, but one issue is the appropriate choice of discretisation of the data for building the BN model. Another issue is the software integration of *OPNET* and *Hugin* to test our framework in a real-time scenario. Work is in progress to address these issues.
2. *Complexity*: The learning algorithms used in our framework are standard and have been optimised for building the BN models, so we can safely conclude that the complexity of our solution is not high.

3. *Speed*: This issue needs to be addressed by conducting further experiments on resource utilisation. However, in the cases which we have considered, the time taken for training and decision making have been found to be quite low (a few milliseconds on a PC having Intel P4 CPU with speed of 3.16 GHz).
4. *Scalability*: To extend this work to large scale networks is an open issue and this will be part of our ongoing work.

This discussion leads us to a promising conclusion that the overall benefits which have been achieved using the BN-based DSS framework, outweigh the costs of computational and implementation complexity. This will form the premise for our future work in this domain.

## 7 Conclusion

This paper has demonstrated that BN are capable of representing dynamic systems, by efficiently modelling unknown and complex relationships between network elements. BN can be used for reasoning, decision making and prediction in the absence of sufficient observable data. We have shown the practicality of BN methodology with the help of a case study related to critical network management function of CAC to guarantee QoS in a heterogeneous IP network. The results demonstrate that BN modelling provides intelligent and automated solutions to improve the functionality of current NMSs. Further work is planned to study the scalability of this methodology to large scale networks. The validation of our simulation-based solution is planned to be performed on a real telecom network. Finally, it would be interesting to compare BN approach to Neural Networks technique in terms of prediction accuracy, model training speed and algorithm complexity.

## Acknowledgement

The authors would like to acknowledge the support of the University of Ulster and IU-ATC for funding this research work through a Vice Chancellors Research Studentship.

## References

1. ITU-T: General overview of NGN, ITU-T Recommendation Y.2001 (2004)
2. Pras, A., et al.: Key research challenges in network management. *IEEE Communications Magazine*, 104–110 (2007)
3. Alpaydin, E.: *Introduction to Machine Learning*. MIT Press, Cambridge (2004)
4. Harrington, D., Presuhn, R., Wijnen, B.: An architecture for describing SNMP management frameworks. RFC 3411, IETF (2002)
5. ITU-T: TMN Management Functions, ITU-T Recommendation M.3400 (2000)
6. Kulkarni, P.G., McClean, S.I., Parr, G.P., Black, M.M.: Deploying MIB data mining for proactive network management. In: 3rd Intl. IEEE Conference on Intelligent Systems, pp. 506–511 (2006)

7. Sohail, S., Khanum, A.: Simplifying network management with fuzzy logic. In: IEEE Intl. Conf. on Communications, pp. 195–201 (2008)
8. Hood, C.S., Ji, C.: Proactive network fault detection. IEEE Transactions on Reliability, 333–341 (1997)
9. Ding, J., Kramer, B., et al.: Predictive fault management in the dynamic environment of IP network. In: IEEE International Workshop on IP Operations and Management, pp. 233–239 (2004)
10. Ekaette, E.U., Far, B.H.: A framework for distributed fault management using intelligent software agents. In: IEEE Canadian Conference on Electrical and Computer Engineering, vol. 2, pp. 797–800 (2003)
11. Hui, T.C.K., Chen-Khong, T.: Adaptive provisioning of differentiated services networks based on reinforcement learning. IEEE Transactions on Systems, Man, and Cybernetics 33(4), 492–501 (2003)
12. Bashar, A., Parr, G.P., McClean, S.I., Scotney, B.W., Nauck, D.: BARD: A novel application of Bayesian reasoning for proactive network management. In: 10th Annual Postgraduate Conference on Telecommunications, Networking and Broadcasting (PGNET 2009), pp. 161–166 (2009)
13. Bashar, A., et al.: Employing Bayesian belief networks for energy efficient network management. In: IEEE National Conference on Communications (NCC) 2010, pp. 1–5 (2010)
14. Qi, J., Wu, F., Li, L., Shu, H.: Artificial intelligence applications in the telecommunications industry. Expert Systems 24, 271–291 (2007)
15. Heckerman, D.: A tutorial on learning with Bayesian Networks. In: Jordan, M. (ed.) Learning in Graphical Models. MIT Press, Cambridge (1999)
16. Spirtes, P., Glymour, C., Scheines, R.: Causation, Prediction, and Search, 2nd edn. MIT Press, Cambridge (2001)
17. Steck, H.: Constrained-based structural learning in Bayesian networks using finite data sets. PhD Thesis, Institut für der Informatik der Technischen Universität München (2001)
18. Jensen, F.: Bayesian Networks and Decision Graphs, 2nd edn. Springer, Heidelberg (2007)
19. Murphy, K.: Dynamic Bayesian networks (2002), <http://people.cs.ubc.ca/murphyk/Papers/dbnchapter.pdf>
20. Korb, K.B., Nicholson, A.E.: Bayesian Artificial Intelligence, 1st edn. CRC Press, Boca Raton (2003)
21. Wright, S.: Admission Control in Multi-Service IP Networks: A Tutorial. IEEE Communications Surveys & Tutorials, 72–86 (2007)
22. OPNET Modeler 16.0, <http://www.opnet.com>
23. Hugin Researcher 7.2, <http://www.hugin.com>
24. Cisco Systems: Network element polling with Cisco active network abstraction (2007), <http://www.ciscosystems.com/>
25. Nam, S.Y., et al.: Measurement-based admission control at edge routers. IEEE/ACM Transactions on Networking 16(2), 410–423 (2008)
26. Nichols, K., et al.: Definition of the Differentiated Services Field (DS Field) in the IPv4 and IPv6 Headers. RFC 2474, IETF (1998)

# Combining Logic and Probabilities for Discovering Mappings between Taxonomies

Rémi Tournaire<sup>1</sup>, Jean-Marc Petit<sup>2</sup>, Marie-Christine Rousset<sup>1</sup>,  
and Alexandre Termier<sup>1</sup>

<sup>1</sup> University of Grenoble, Laboratory of Informatics of Grenoble UMR 5217,  
681, rue de la Passerelle, BP72, 38402 St-Martin d'Hères Cedex, France  
Remi.Tournaire@imag.fr

<sup>2</sup> INSA Lyon, LIRIS UMR 5205, 69621 Villeurbanne Cedex, France

**Abstract.** In this paper, we investigate a principled approach for defining and discovering *probabilistic mappings* between two taxonomies. First, we compare two ways of modeling probabilistic mappings which are compatible with the logical constraints declared in each taxonomy. Then we describe a *generate and test* algorithm which minimizes the number of calls to the probability estimator for determining those mappings whose probability exceeds a certain threshold. Finally, we provide an experimental analysis of this approach.

## 1 Introduction

The decentralized nature of the development of Web data management systems makes inevitable the independent construction of a large amount of personalized taxonomies used for annotating data and resources at Web scale. Taxonomies are hierarchical structures appropriate for data categorization and semantic annotation of resources. They play a prominent role in the Semantic Web since they are central components of OWL [8] or RDF(S) [19] ontologies. A taxonomy constrains the vocabulary used to express metadata or semantic annotations to be classes that are related by structural relationships. Taxonomies are easy to create and understand by humans while being machine interpretable and processable thanks to a formal logical semantics supporting reasoning capabilities.

In this setting, establishing *semantic mappings* between taxonomies is the key to enable collaborative exchange of semantic data. Manually finding such mappings is clearly not possible at the Web scale. Therefore, the automatic discovery of semantic mappings is the bottleneck for scalability purposes.

Many techniques and prototypes have been developed to suggest candidate mappings between several knowledge representations including taxonomies, ontologies or schemas (see [25,14] for surveys). Most of the existing approaches rely on evaluating the degree of similarity between the elements (e.g., classes, properties, instances) of one ontology and the elements of another ontology. Many different similarity measures are proposed and often combined. Most of them are based on several syntactic, linguistic or structural criteria to measure the proximity of the terms used to denote the classes and/or their properties within the

ontology. Some of them exploit characteristics of the data declared as instances of the classes (e.g. [12]).

As a result, most of the existing matching systems return for every candidate pair of elements a coefficient in the range  $[0,1]$  which denotes the strength of the semantic correspondence between those two elements [15,24,4]. A threshold is then used for keeping as *valid mappings* those pairs of elements for which the coefficient of similarity is greater than the threshold. Since most of the approaches are based on similarity functions that are symmetric, the mappings that are returned with high similarity scores are interpreted as *equivalence mappings*. Few approaches [17,18] handle *inclusion mappings* between classes. Yearly international evaluation campaigns<sup>1</sup> are organized to compare matching systems on different benchmarks, in terms of quality (recall and precision) of the mappings they return. Except until very recently, only equivalence mappings have been considered in the OAEI campaigns.

Our first claim is that *inclusion* mappings between classes of two pre-existing taxonomies are more likely to exist than *equivalence* mappings. When taxonomies are used as query interfaces between users and data, inclusion mappings between taxonomies can be used for query reformulation exactly like the subclass relationship within a taxonomy. For instance, a mapping  $Opera \sqsubseteq Vocal$  between the class *Opera* of a taxonomy and the class *Vocal* of a second taxonomy may be used to find additional answers to a query asking data about *Vocal* by returning data categorized in the class *Opera* in the first taxonomy.

In contrast with logical approaches (e.g., [17]) for (inclusion) mapping discovery, we also claim that *uncertainty* is intrinsic to mapping discovery. Therefore, we advocate to consider *inclusion mappings with a probabilistic semantics*. Like the similarity scores, the probability coefficients can be compared to a threshold for filtering mappings. In addition, they can be the basis of a probabilistic reasoning and a probabilistic query answering through mapped taxonomies.

It is important to emphasize here that the similarity coefficients returned by most of the existing ontology or schema matching systems cannot be interpreted as *probabilities* of the associated mappings. The reason is that they do not take into account possible logical implications between mappings, which can be inferred from the inclusion axioms declared between classes within each ontology. Interpreting similarities between classes as probabilities of the corresponding mappings requires that the similarity between any subclass of a given class  $A_1$  and any superclass of a given class  $A_2$  is greater than the similarity between  $A_1$  and  $A_2$ . Up to our knowledge, this monotony property is not satisfied in any of the existing similarity models.

In this paper, we propose an algorithm for automatic discovery of *probabilistic mappings* between taxonomies, which respects the above monotony property. First, we investigate and compare two ways of modeling probabilistic mappings which are compatible with the logical constraints declared in each taxonomy. In those two probabilistic models, the probability of a mapping relies on the joint probability distribution of the involved classes. They differ on the property of

---

<sup>1</sup> E.g., OAEI <http://oaei.ontologymatching.org/2009/>

*monotony* of the corresponding probability function with respect to the logical implication. Based on the above probabilistic setting, we have designed, implemented and experimented a *generate and test* algorithm called ProbaMap for discovering the mappings whose probability is greater than a given threshold. In this algorithm, the monotony of the probability function is exploited for avoiding the probability estimation of as many mappings as possible. The paper is organized as follows. Section 2 presents the formal background and states the problem considered in this paper. Section 3 is dedicated to the definition and computation of mapping probabilities. In Section 4, we present the ProbaMap algorithm which discovers mappings with high probabilities (i.e., greater than a threshold). Section 5 surveys the quantitative and qualitative experiments that we have done. Finally, in Section 6, we compare our approach to existing works and we conclude.

## 2 Formal Background

We first define taxonomies as a graphical notation and its interpretation in standard first-order-logic semantics, on which the inheritance of instances is grounded. Then, we define *mappings* between taxonomies as inclusion statements between classes of two different taxonomies. Finally, we set the problem statement of matching taxonomies that we consider in this paper.

### Taxonomies: classes and instances

Given a vocabulary  $\mathcal{V}$  denoting a set of classes, a *taxonomy*  $\mathcal{T}_y$  is a Directed Acyclic Graph (DAG) where each node is labelled with a distinct *class* name of  $\mathcal{V}$ , and each arc between a node labelled with  $C$  and a node labelled by  $D$  represents a *specialization relation* between the classes  $C$  and  $D$ .

Each class in a taxonomy can be associated with a set of *instances* which have an *identifier* and a content *description*. In the following, we will abusively speak of the instance  $i$  to refer to the instance identified by  $i$ . Figure 1 shows

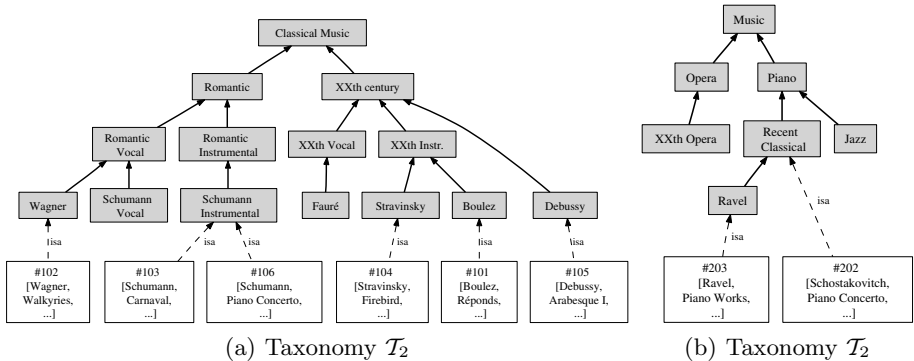


Fig. 1. 2 Taxonomies and associated instances

two samples of taxonomies related to the Music domain. Bold arrows are used for representing specialization relations between classes, and dashed arrows for membership relation between instances and classes. In both taxonomies, some instances, with description denoted between brackets, are associated to classes. For example, #102 is an instance identifier and [Wagner, Walkyries, ...] its associated description.

The instances that are in the scope of our data model can be web pages (which content description is a set of words) identified by their URLs, RDF resources (which content description is a set of RDF triples) identified by URIs, or audio or video files identified by a signature and whose content description may be attribute-value metadata that can be extracted from those files. Taxonomies have a logical semantics which provides the basis to define formally the extension of a class as the set of instances that are declared or can be *inferred* for that class.

### Logical semantics

There are several graphical or textual notations for expressing the specialization relation between a class  $C$  and a class  $D$  in a taxonomy. For example, in RDF(S) [19] which is the first standard of the W3C concerning the Semantic Web, it is denoted by  $(C \text{ rdfs:subclassOf } D)$ . It corresponds to the inclusion statement  $C \sqsubseteq D$  in the description logics notation.

Similarly, a membership statement denoted by an *isa* arc from an instance  $i$  to a class  $C$  corresponds in the RDF(S) notation to  $(i \text{ rdf:type } C)$ , and to  $C(i)$  in the usual notation of description logics.

All those notations have a standard model-theoretic logical semantics based on interpreting classes as sets: an *interpretation*  $\mathcal{I}$  consists of a non empty domain of interpretation  $\Delta^{\mathcal{I}}$  and a function  $\cdot^{\mathcal{I}}$  that interprets each class as a non empty subset of  $\Delta^{\mathcal{I}}$ , and each instance identifier as an element of  $\Delta^{\mathcal{I}}$ . The classes declared in a taxonomy are interpreted as non empty subsets because they are object containers. According to the *unique name assumption*, two distinct identifiers  $a$  and  $b$  verify  $(a^{\mathcal{I}} \neq b^{\mathcal{I}})$  in any interpretation  $\mathcal{I}$ .

$\mathcal{I}$  is a *model* of a taxonomy  $\mathcal{T}$  if:

- for every inclusion statement  $E \sqsubseteq F$  of  $\mathcal{T}$ :  $E^{\mathcal{I}} \subseteq F^{\mathcal{I}}$ ,
- for every membership statement  $C(a)$  of  $\mathcal{T}$ :  $a^{\mathcal{I}} \in C^{\mathcal{I}}$ .

An inclusion  $G \sqsubseteq H$  is *inferred* by a taxonomy  $\mathcal{T}$  (denoted by  $\mathcal{T} \models G \sqsubseteq H$ ) iff in every model  $\mathcal{I}$  of  $\mathcal{T}$ ,  $G^{\mathcal{I}} \subseteq H^{\mathcal{I}}$ . A membership  $C(e)$  is *inferred* by  $\mathcal{T}$  (denoted by  $\mathcal{T} \models C(e)$ ) iff in every model  $\mathcal{I}$  of  $\mathcal{T}$ ,  $e^{\mathcal{I}} \in C^{\mathcal{I}}$ .

Let  $\mathcal{D}$  be the set of the instances associated with a taxonomy  $\mathcal{T}$ . The *extension* of a class  $C$  in  $\mathcal{T}$ , denoted by  $Ext(C, \mathcal{T})$ , is the set of instances for which it can be inferred from the membership and inclusion statements declared in the taxonomy that they are instances of  $C$ :  $Ext(C, \mathcal{T}) = \{d \in \mathcal{D} / \mathcal{T} \models C(d)\}$ .

### Mappings

The mappings that we consider are inclusion statements involving classes of two different taxonomies  $\mathcal{T}_1$  and  $\mathcal{T}_2$ . To avoid ambiguity and without loss of generality, we consider that each taxonomy has its own vocabulary: by convention

we index the names of the classes by the index of the taxonomy to which they belong. Mappings between  $\mathcal{T}_1$  and  $\mathcal{T}_2$  are consequently of the form  $A_1 \sqsubseteq B_2$  or  $A_2 \sqsubseteq B_1$ . For a mapping  $m$  of the form  $A_i \sqsubseteq B_j$ , its left-hand side  $A_i$  will be denoted  $lhs(m)$  and its right-hand side will be denoted  $rhs(m)$ .

A mapping  $A_i \sqsubseteq B_j$  has the same meaning as a specialization relation between the classes  $A_i$  and  $B_j$ , and thus is interpreted in logic in the same way, as a set inclusion. The logical entailment between classes extends to logical entailment between mappings as follows.

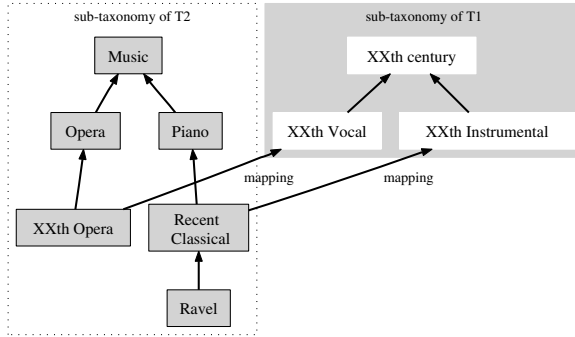


Fig. 2. 2 mappings between  $\mathcal{T}_1$  and  $\mathcal{T}_2$

**Definition 1 (Entailment between mappings).** Let  $\mathcal{T}_i$  and  $\mathcal{T}_j$  be two taxonomies. Let  $m$  and  $m'$  be two mappings between  $\mathcal{T}_i$  and  $\mathcal{T}_j$ :  $m$  entails  $m'$  (denoted  $m \preceq m'$ ) iff every model of  $\mathcal{T}_i, \mathcal{T}_j$  and  $m$  is also a model of  $m'$ .

It is straightforward to show that  $\preceq$  is a (partial) order relation on the set of mappings between the two taxonomies  $\mathcal{T}_i$  and  $\mathcal{T}_j$ . If  $m \preceq m'$ , we will say that  $m$  is more specific than  $m'$  (also that  $m$  is an implicant of  $m'$ ) and that  $m'$  is more general than  $m$  (also that  $m'$  is an implicate of  $m$ ).

The following proposition characterizes the logical entailment between mappings in function of the logical entailment between the classes of their left hand sides and right hand sides.

**Proposition 1.** Let  $m$  and  $m'$  be two mappings between two taxonomies. Let  $\mathcal{T}_i$  be the taxonomy of  $lhs(m)$ ,  $\mathcal{T}_j$  the taxonomy of  $rhs(m)$ .

$m \preceq m'$  iff

- $lhs(m)$  and  $lhs(m')$  are classes of the same taxonomy  $\mathcal{T}_i$ , and
- $\mathcal{T}_i \models lhs(m') \sqsubseteq lhs(m)$  and  $\mathcal{T}_j \models rhs(m) \sqsubseteq rhs(m')$

For example, two mappings between taxonomies  $\mathcal{T}_1$  and  $\mathcal{T}_2$  of Figure 1 are illustrated in Figure 2. The mapping  $XXth\ Opera_2 \sqsubseteq XXth\ Vocal_1$  is more specific than the mapping  $XXth\ Opera_2 \sqsubseteq XXth\ Century_1$ , and the mapping  $RecentClassical_2 \sqsubseteq XXth\ Instrumental_1$  is more specific than the mapping  $Ravel_2 \sqsubseteq XXth\ Century_1$ .



### 3 Mapping Probabilities: Models and Estimation

We consider two probabilistic models for modeling uncertain mappings. They are both based on the discrete probability measure defined on subsets of the sample set representing the set of all possible instances of the two taxonomies. From now on, we will denote  $Pr(E)$  the probability for an instance to be an element of the subset  $E$ .

The first model defines the probability of a mapping  $A_i \sqsubseteq B_j$  as the conditional probability for an instance to be an instance of  $B_j$  knowing that it is an instance of  $A_i$ . It is the natural way to extend the logical semantics of entailment to probabilities.

The second model comes directly from viewing classes as subsets of the sample space: the probability of  $A_i \sqsubseteq B_j$  is the probability for an element to belong to the set  $\overline{A_i} \cup B_j$ , where  $\overline{A_i}$  denotes the complement set of  $A_i$  in the sample set. Both models are described below.

**Definition 2 (Two probabilities for a mapping).** *Let  $m$  be a mapping of the form  $A_i \sqsubseteq B_j$ .*

*-Its conditional probability, denoted  $P_c(m)$ , is defined as  $P_c(m) = Pr(B_j|A_i)$ .*

*-Its union set probability, denoted  $P_u(m)$ , is defined as  $P_u(m) = Pr(\overline{A_i} \cup B_j)$ .*

Proposition 2 states the main (comparative) properties of those two probabilistic models. They both meet the logical semantics for mappings that are certain, and they can both be expressed using joint probabilities.

**Proposition 2.** *Let  $m$  be a mapping between two taxonomies  $\mathcal{T}_i$  and  $\mathcal{T}_j$ . The following properties hold:*

1.  $P_u(m) \geq P_c(m)$ .
2. If  $m$  is a certain mapping,  $P_c(m) = P_u(m) = 1$
3.  $P_u(m) = 1 + Pr(lhs(m) \cap rhs(m)) - Pr(lhs(m))$
4.  $P_c(m) = \frac{Pr(lhs(m) \cap rhs(m))}{Pr(lhs(m))}$

They differ on the monotony property w.r.t the (partial) order  $\preceq$  corresponding to logical implication (cf. Definition 1):  $P_u$  is monotonous whereas  $P_c$  verifies a property of *weak* monotony only:

**Theorem 1 (Property of monotony).** *Let  $m$  and  $m'$  two mappings.*

1. If  $m \preceq m'$  then  $P_u(m) \leq P_u(m')$
2. If  $m \preceq m'$  and  $lhs(m) = lhs(m')$ ,  $P_c(m) \leq P_c(m')$

The proof results from Proposition 1 and Proposition 2 which relate mappings with the classes of their left hand sides and right hand sides for logical entailment and probabilities respectively, and from considering (declared or inherited) inclusions of classes within each taxonomy as statements whose probability is equal to 1.

As shown in Proposition 2, the computation of  $P_u(m)$  and  $P_c(m)$  relies on computing the set probability  $Pr(lhs(m))$  and the joint set probability  $Pr(lhs(m) \cap rhs(m))$ .

$\cap rhs(m)$ ). Those values are unknown and must be estimated. For doing so, we follow the Bayesian approach to statistics [9]: we model those (unknown) parameters as continuous random variables, and we use *observations* to infer their *posterior* distribution from their *prior* distribution. This is summarized in Definition 3.

**Definition 3 (Bayesian estimator of  $Pr(E)$ ).** *Let  $E$  be a subset of the sample set  $\Omega$ . Let  $\mathcal{O}$  be a sample of observed elements for which it is known whether they belong or not to  $E$ . The Bayesian estimator of  $Pr(E)$ , denoted  $\widehat{Pr}(E)$ , is the expected value of the posterior distribution of  $Pr(E)$  knowing the observations on the membership to  $E$  of each element in  $\mathcal{O}$ , and setting the prior probability of a random set to  $\frac{1}{2}$ , and of the intersection of two random sets to  $\frac{1}{4}$ .*

Setting the prior probabilities to  $\frac{1}{2}$  and  $\frac{1}{4}$  depending on whether  $E$  is a class or a conjunction of classes corresponds to the uniform distribution of instances among the classes. The following theorem provides a simple way to compute the Bayesian estimations  $\widehat{P}_u(m)$  and  $\widehat{P}_c(m)$  of the two probabilities  $P_u(m)$  and  $P_c(m)$  defined in Definition 2. It is a straightforward consequence of a basic theorem in probability theory (Theorem 1, page 160, [9]), stating that if the prior distribution of the random variable modeling  $Pr(E)$  is a *Beta distribution* of parameters  $\alpha$  and  $\beta$ , then its posterior distribution is also a Beta distribution the parameters of which are:  $\alpha + |Ext(E, \mathcal{O})|$  and  $\beta + |\mathcal{O}|$ , where  $Ext(E, \mathcal{O})$  is the set of observed instances of  $\mathcal{O}$  that are recognized to belongs to  $E$ .

**Theorem 2 (Estimation of probabilities)**

*Let  $m : C_i \sqsubseteq D_j$  be a mapping between two taxonomies  $\mathcal{T}_i$  and  $\mathcal{T}_j$ . Let  $\mathcal{O}$  be the union of instances observed in  $\mathcal{T}_i$  and  $\mathcal{T}_j$ . Let  $N = |\mathcal{O}|$ ,  $N_i = |Ext(C_i, \mathcal{O})|$ ,  $N_j = |Ext(D_j, \mathcal{O})|$  and  $N_{ij} = |Ext(C_i \cap D_j, \mathcal{O})|$ .*

$$\widehat{P}_u(m) = 1 + \frac{1+N_{ij}}{4+N} - \frac{1+N_i}{2+N} \qquad \widehat{P}_c(m) = \frac{1+N_{ij}}{4+N} \times \frac{2+N}{1+N_i}$$

Depending on the way the taxonomies are populated (manually or automatically), it is not always possible to obtain  $N_{ij}$  simply by counting the instances that are common to the two classes involved in the mapping. If the taxonomies are populated manually and independently by different users, it is indeed likely that the intersection of the two taxonomies contains very few instances or even no instance at all. In that case, we apply existing *automatic classifiers* (e.g., Naive Bayes learning, decision trees, SVM) in order to compute  $Ext(C_i \cap D_j, \mathcal{O})$ , by following the same approach as [12] for training them on the description of the available instances in each taxonomy.

**4 The ProbaMap Algorithm**

Given two taxonomies  $\mathcal{T}_i$  and  $\mathcal{T}_j$  (and their associated instances), let  $\mathcal{M}(\mathcal{T}_i, \mathcal{T}_j)$  be the set of all mappings from  $\mathcal{T}_i$  to  $\mathcal{T}_j$  (i.e., of the form  $C_i \sqsubseteq D_j$ ). The ProbaMap algorithm determines all mappings  $m$  of  $\mathcal{M}(\mathcal{T}_i, \mathcal{T}_j)$  verifying a probabilistic-based criterion of validity that will be denoted by  $\widehat{P}(m) \geq S$ .

$\widehat{P}(m) \geq S$  is a parameter in the algorithm, which can be one of the three following validity criteria, where  $S_u$  and  $S_c$  are two thresholds in  $[0, 1]$ :

- Validity criterion 1:  $\widehat{P}_u(m) \geq S_u$
- Validity criterion 2:  $\widehat{P}_c(m) \geq S_c$
- Validity criterion 3:  $\widehat{P}_c(m) \geq S_c$  and  $\widehat{P}_u(m) \geq S_u$ .

### Candidate mapping generation

The principle of ProbaMap algorithm is to generate mappings from the two sets of classes in the two taxonomies ordered according to a *topological sort* [6]. Namely, the nested loops (Line 2) in Algorithm 1 generate all the mappings  $C_i \sqsubseteq D_j$  by enumerating the classes  $C_i$  of  $\mathcal{T}_i$  following a *reverse* topological order and the classes  $D_j$  of  $\mathcal{T}_j$  following a *direct* topological order. The following proposition is a corollary of Proposition 1.

**Proposition 3.** *Let  $\mathcal{T}_i$  and  $\mathcal{T}_j$  two taxonomies.*

*Let  $ReverseTopo(\mathcal{T}_i)$  be the sequence of classes of  $\mathcal{T}_i$  resulting from a reverse topological sort of  $\mathcal{T}_i$ . Let  $Topo(\mathcal{T}_j)$  be the sequence of classes of  $\mathcal{T}_j$  resulting from a topological sort of  $\mathcal{T}_j$ . Let  $m : C_i \sqsubseteq D_j$  and  $m' : C'_i \sqsubseteq D'_j$  two mappings from  $\mathcal{T}_i$  to  $\mathcal{T}_j$ . If  $m'$  is an implicant of  $m$  (i.e.,  $m' \preceq m$ ), then  $C_i$  is before  $C'_i$  in  $ReverseTopo(\mathcal{T}_i)$  or  $C_i = C'_i$  and  $D_j$  is before  $D'_j$  in  $Topo(\mathcal{T}_j)$ .*

### Pruning the candidate mappings to test

Based on the monotony property of the probability function  $P_u$  (Theorem 1), every mapping  $m'$  implicant of a mapping  $m$  such that  $P_u(m) < S_u$  verifies  $P_u(m') < S_u$ . Therefore, in ProbaMap, if the validity criterion involves  $\widehat{P}_u$ , we prune the probability estimation of all the implicants of every  $m$  such that  $\widehat{P}_u(m) < S_u$ . We shall use the notation  $Implicants(m)$  to denote the set of all mappings that are implicants of  $m$ . Similarly, based on the property of weak monotony of the probability function  $P_c$  (Theorem 1), if the validity criterion involves  $\widehat{P}_c$ , when a tested candidate mapping  $m$  is such that  $\widehat{P}_c(m) < S_c$  we prune the probability estimation of all the implicants of  $m$  having the same left-hand side as  $m$ . We shall denote this set:  $Implicants_c(m)$ . Based on Proposition 1,  $Implicants(m)$  and  $Implicants_c(m)$  can be generated from  $\mathcal{T}_i$  and  $\mathcal{T}_j$ .

Based on the order in which the mappings are generated, Proposition 3 shows that the validity test in Line 5 of the algorithm 1 maximizes the number of pruning. The resulting ProbaMap algorithm is described in Algorithm 1, in which:

- $\widehat{P}(m) \geq S$  in Line 6 denotes a generic validity criterion that can be instantiated either by  $\widehat{P}_u \geq S_u$ , or by  $\widehat{P}_c \geq S_c$ , or by  $(\widehat{P}_c \geq S_c$  and  $\widehat{P}_u \geq S_u)$ .
- In the case where the validity criteria involves  $\widehat{P}_c$ ,  $Implicants(m)$  in Line 9 must be replaced by  $Implicants_c(m)$ .
- In Line 4,  $ReverseTopo_i$  and  $Topo_j$  denote the respective sequences  $ReverseTopo(\mathcal{T}_i)$  and  $Topo(\mathcal{T}_j)$ .  $ReverseTopo_i[k]$  (resp.  $Topo_j[l]$ ) denotes the class of  $\mathcal{T}_i$  (resp.  $\mathcal{T}_j$ ) ranked  $k$  (resp.  $l$ ) in the sequence.

Algorithm 1 returns mappings directed from  $\mathcal{T}_i$  to  $\mathcal{T}_j$ . In order to obtain *all* valid mappings, it must be applied again by swapping its inputs  $\mathcal{T}_i$  and  $\mathcal{T}_j$ .

---

**Algorithm 1.** ProbaMap

---

**Require:**  $\mathcal{T}_i, \mathcal{T}_j, \text{threshold } S$ **Ensure:** return  $\{m \in \mathcal{M}(\mathcal{T}_i, \mathcal{T}_j) / \hat{P}(m) \geq S\}$ 

```

1:  $M_{Val} \leftarrow \emptyset, M_{NVal} \leftarrow \emptyset$ 
2: for  $k = 1$  to  $|\mathcal{T}_i|$  do
3:   for  $l = 1$  to  $|\mathcal{T}_j|$  do
4:     let  $m = \text{ReverseTopo}_i[k] \sqsubseteq \text{Topo}_j[l]$ 
5:     if  $m \notin M_{NVal}$  then
6:       if  $\hat{P}(m) \geq S$  then
7:          $M_{Val} \leftarrow M_{Val} \cup \{m\}$ 
8:       else
9:          $M_{NVal} \leftarrow M_{NVal} \cup \text{Implicants}(m)$ 
10: return  $M_{Val}$ 

```

---

## 5 Experiments

In this section, we evaluate the quantitative and qualitative performances of ProbaMap (Algorithm 1) on large real-world taxonomies populated with instances. We focus our experiments on the Internet directories<sup>2</sup> from Yahoo! and Google (actually corresponding to Dmoz). This allows us to compare our approach to the SBI algorithm of Ichise et al. [20,21], which is dedicated to the discovery of mappings between Internet directories. Internet directories are huge trees of categories, which can be seen as taxonomies, categories being the classes. Each category contains a set of links (i.e. URLs to web sites), which can be seen as the instances of the class. Each link comes with a small text summary, whose words can be seen as instance attributes for classification.

Our datasets are corresponding locations in the Yahoo! and Google directories, that have also been used in the experiments of [20,21]:

- Yahoo! : Recreation / Automotive & Google : Recreation / Autos
- Yahoo! : Computers\_and\_Internet/Software & Google : Computers/Software
- Yahoo! : Arts / Visual\_Arts / Photography & Google : Arts / Photography

The data from the directories was collected in the beginning of 2010, so is slightly different from the data of [21] and [20] which was collected in Fall 2001.

Table 1 shows for each dataset the number of classes and instances in each class, and the number of instances shared between the Yahoo! and the Google directories. Two instances are considered shared if they correspond to the same URL in both directories. For a fair comparison, we have implemented both ProbaMap and the SBI algorithm in Java.

The goal of our experiments is to compare the quality of Internet directories alignment for ProbaMap and SBI.

During the learning step, ProbaMap and SBI receive as training set a subset of the shared instances with their correct category in each of the directories. The test set is the remaining of the shared instances. When adding classification to

---

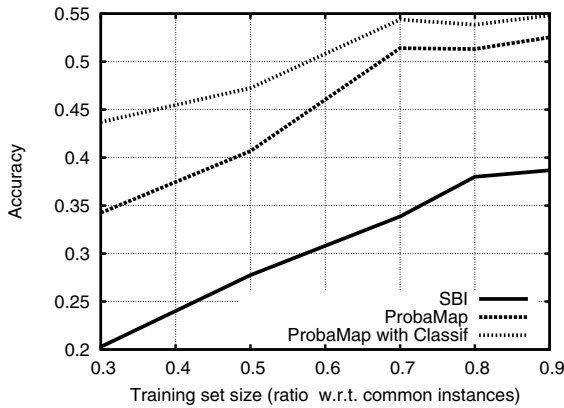
<sup>2</sup> [dir.yahoo.com](http://dir.yahoo.com), [www.dmoz.org](http://www.dmoz.org)

**Table 1.** Statistics on data collected from subdirectories on Yahoo! and Google

	Yahoo!		Google		shared instances
	classes	instances	classes	instances	
<b>Autos</b>	947	4406	967	6425	837
<b>Software</b>	323	2390	2395	30140	572
<b>Photography</b>	168	1851	321	3852	286

ProbaMap, the training set is extended with all the non shared instances. The classification is performed using the SVM implementation SMO[16] in Weka [27], where the classification attributes for an instance are the words of its summary.

We computed the accuracy (standard criteria used in [21]) of mapping prediction on the test set, and conducted a ten-fold cross validation. The results when varying the size of the training set are shown in Figure 3 for the Autos dataset.

**Fig. 3.** Accuracy for alignment of Autos subdirectories (Yahoo to Google)

ProbaMap with or without classification significantly outperforms SBI, with a 15% better accuracy for ProbaMap alone and 15-25% better accuracy for ProbaMap with classification. Note that when using classification, the accuracy of ProbaMap with the smallest training set size (30% of shared instances) is better than the accuracy of SBI with a training set containing 90% of shared instances. We obtain similar results (and coherent with those in [21]) on Software and Photography subdirectories, with the same order for the three methods.

These results show that the probabilistic mapping discovery method that we propose gives excellent results on real-world datasets, and can take advantage of classification techniques to compensate small training set sizes. This is an important quality for real world taxonomies built by different people, that are unlikely to have many instances in common.

Thanks to the monotony property of our probabilistic mapping model (see Section 4), ProbaMap can prune large parts of the mapping search space and handle the alignment of very large taxonomies. We successfully conducted

scalability experiments on very large taxonomies from the OAEI<sup>3</sup> contest (directory dataset). We have compensated the lack of instances available for those taxonomies by automatically populating the classes with WordNet synsets. The taxonomies to align have 6628 and 2857 classes, leading to more than 15 millions potential mappings. Up to our knowledge, very few OAEI participants have taken the whole directory dataset as input, but instead a splitted version of it into branches which was provided by the OAEI organizers.

## 6 Related Work and Conclusion

As outlined in the introduction, semantic mappings are the glue for data integration systems. A wide range of methods of schema/ontology matching have been developed both in the database and the semantic web communities [14]. One of the principles widely exploited is terminological comparison of the labels of classes with string-based similarities or lexicon-based similarities (like WordNet) (e.g., TaxoMap [18], H-MATCH [4]). Another widely used principle is structure comparison between labeled graphs representing ontologies (e.g., OLA [15]). In fact, most of the existing matchers combine these two approaches in different ways (e.g., COMA++ [1] and COMA [10], Cupid [24], H-MATCH [4]). Other approaches have been investigated with machine learning techniques using a corpus of schema matches (e.g., [23]), or a corpus of labelled instances (e.g., LSD [11], SemInt [22], GLUE [12], FCA-merge [26], SBI-NB[21]). SBI[20] computes the degree of agreement of each couple of classes based on instances statistics, but without machine learning. It is standard practice for ontology and schema matchers to associate numbers with the candidate mappings they propose. However, those numbers do not have a probabilistic meaning and are just used for ranking. In contrast, our approach promotes a probabilistic semantics for mappings and provides a method to compute mapping probabilities based on the descriptions of instances from in each ontology. It is important to note that even if we use similar classification techniques as [12], we use them for computing true probabilities and not similarity coefficients.

The most distinguishing feature of our approach is that it bridges the gap between logic and probabilities by providing probabilistic models that are consistent with the logical semantics underlying ontology languages. Therefore, our approach generalizes existing works based on algebraic or logical representation of mappings as a basis for reasoning (e.g., S-Match [17], Clío [5]). The mappings returned by ProbaMap can be exploited for mapping validation by probabilistic reasoning in the line of what is proposed in [3]. More generally, our approach is complementary of the recent work that has been flourishing on probabilistic databases [2,7]. It fits into the general framework set in [13] for handling uncertainty in data integration, for which it provides an effective way for computing mapping probabilities.

The experiments that we have conducted on both real-world and controlled data have shown the feasibility and the scalability of our approach.

---

<sup>3</sup> <http://oaei.ontologymatching.org>

## References

1. Aumueller, D., Do, H.H., Massmann, S., Rahm, E.: Schema and ontology matching with COMA++. In: SIGMOD 2005, ACM, New York (2005)
2. Benjelloun, O., Sarma, A.D., Halevy, A.Y., Widom, J.: ULDBs: Databases with uncertainty and lineage. In: VLDB (2006)
3. Castano, S., Ferrara, A., Lorusso, D., N ath, T.H., M oller, R.: Mapping validation by probabilistic reasoning. In: Bechhofer, S., Hauswirth, M., Hoffmann, J., Koubarakis, M. (eds.) ESWC 2008. LNCS, vol. 5021, pp. 170–184. Springer, Heidelberg (2008)
4. Castano, S., Ferrara, A., Montanelli, S.: H-MATCH: an algorithm for dynamically matching ontologies in peer-based systems. In: SWDB (2003)
5. Chiticariu, L., Hern andez, M.A., Kolaitis, P.G., Popa, L.: Semi-automatic schema integration in clio. In: VLDB (2007)
6. Cormen, T.H., Leiserson, C.E., Rivest, R.L., Stein, C.: Introduction to Algorithms, 2nd edn. The MIT Press, Cambridge (2001)
7. Dalvi, N.N., Suciu, D.: Answering queries from statistics and probabilistic views. In: VLDB (2005)
8. Dean, M., Schreiber, G.: OWL web ontology language reference. W3C recommendation, W3C (2004)
9. Degroot, M.H.: Optimal Statistical Decision. Wiley Classics Library (2004)
10. Do, H., Rahm, E.: COMA - a system for flexible combination of schema matching approaches. In: VLDB (2002)
11. Doan, A., Domingos, P., Levy, A.Y.: Learning mappings between data schemas. In: Proceedings of the AAAI 2000 Workshop on Learning Statistical Models from Relational Data (2000)
12. Doan, A., Madhavan, J., Domingos, P., Halevy, A.Y.: Learning to map between ontologies on the semantic web. In: WWW (2002)
13. Dong, X.L., Halevy, A.Y., Yu, C.: Data integration with uncertainty. VLDB Journal (2007)
14. Euzenat, J., Shvaiko, P.: Ontology matching. Springer, Heidelberg (2007)
15. Euzenat, J., Valtchev, P.: Similarity-based ontology alignment in OWL-lite. In: ECAI (2004)
16. Flake, G.W., Lawrence, S.: Efficient SVM regression training with SMO. Machine Learning (2002)
17. Giunchiglia, F., Shvaiko, P., Yatskevich, M.: S-Match: an algorithm and an implementation of semantic matching. In: Bussler, C.J., Davies, J., Fensel, D., Studer, R. (eds.) ESWC 2004. LNCS, vol. 3053, pp. 61–75. Springer, Heidelberg (2004)
18. Hamdi, F., Zargayouna, H., Safar, B., Reynaud, C.: TaxoMap in the OAEI 2008 alignment contest. In: OAEI 2008 Campaign - Int. Workshop on Ontology Matching (2008)
19. Hayes, P. (ed.): RDF Semantics. World Wide Web Consortium (2004)
20. Ichise, R., Takeda, H., Honiden, S.: Integrating multiple internet directories by instance-based learning. In: IJCAI, vol. 18 (2003)
21. Ichise, R., Hamasaki, M., Takeda, H.: Discovering relationships among catalogs. In: Suzuki, E., Arikawa, S. (eds.) DS 2004. LNCS (LNAI), vol. 3245. Springer, Heidelberg (2004)
22. Li, W.S., Clifton, C.: Semint: a tool for identifying attribute correspondences in heterogeneous databases using neural networks. Data Knowl. Eng. 33(1) (2000)

23. Madhavan, J., Bernstein, P.A., Doan, A., Halevy, A.: Corpus-based schema matching. In: International Conference on Data Engineering (2005)
24. Madhavan, J., Bernstein, P.A., Rahm, E.: Generic schema matching with cupid. VLDB Journal (2001)
25. Rahm, E., Bernstein, P.A.: A survey of approaches to automatic schema matching. VLDB Journal (2001)
26. Stumme, G., Maedche, A.: FCA-MERGE: Bottom-Up Merging of Ontologies. In: Proc. of the 17th International Joint Conference on Artificial Intelligence (2001)
27. Witten, I.H., Frank, E.: Data Mining: Practical Machine Learning Tools and Techniques, 2nd edn. Morgan Kaufmann, San Francisco (2005)



# An Ontology-Based Semantic Web Service Space Organization and Management Model

Kun Yang<sup>1,2,\*</sup> and Zhongzhi Shi<sup>1</sup>

<sup>1</sup> Key Lab of Intelligent Information Processing, Institute of Computing Technology,  
Chinese Academy of Sciences, 100190 Beijing, China

<sup>2</sup> Graduate University of the Chinese Academy of Sciences, 100049 Beijing, China  
yangkun@ics.ict.ac.cn,  
shizz@ict.ac.cn

**Abstract.** Semantic Web services facilitate the handling of Web services automatically and precisely with elaborate semantic descriptions for them, which at the same time makes the interactions between them more complicated and time-consuming, and if the number of Web services involved in an task is vast, the time cost is unacceptable. This paper introduce an ontology-based multi-dimension model which organize and manage the Semantic Web services space from different aspects, and different dimensions in it can cooperate each other to locate Web Services needed quickly. This model can improve efficiency by restricting Web Services to be handled in a smaller space and reducing the time of invalid handling to those irrelevant. The implementation mechanism of the model and how to use it to support Web Service discovery and composition are both discussed.

**Keywords:** Ontology-based, semantic web service, multi-dimensional model, service discovery, service composition.

## 1 Introduction

In order to facilitate the automated and precise handling of Web Services, people begin to describe them at semantic level to make them processable entities for computers. Some representative ontology-based semantic description languages, such as OWL-S[1], WSMO[2], METEOR-S[3], are used to describe the static semantic features of Web Services, and some mature process models, such as state machine, Petri net and process algebra, are introduced to describe the dynamic features of them. All these semantic descriptions describe Web services comprehensively and exactly, which make the automatic Web service discovery, selection and execution possible. While at the same time, they also make the interaction between Web Services more complicated and time-consuming. With

---

\* This work is supported by the National Science Foundation of China (60775035), 863 National High-Tech Program (No.2007AA01Z132), National Basic Research Priorities Programme (No. 2003CB317004, 2007CB311004) and National Science and Technology Support Plan (No.2006BAC08B06).

the exponentially increasing of Web services and the interactions between them becoming more complicated, the problem of efficiency becomes more urgent and meaningful than ever before, and many researches have been studied around it. From the aspect of organization of Web Services space, some techniques are adopted. UDDI provides several service classification systems including North America Industry classification system, universal standard products and service classification system, but the standards of these classification system are inconsistent and coarse, and they can only classify service based on their areas or locations rather than their functions. [4] organizes Web Services based on their output, it builds a inverted index for the outputs of all registered services in service library and maintains a services list for each output to record all the services that produce that output. [5][6] cluster similar services together based on their specific single feature.

All techniques introduced above only consider a single attribute of Web Services, and that is not enough, because people often consider several aspects of Web Services during their Web Service discovery, including their locations, their functions, their qualities and so on. Organizing Web Service space based on attributes in various aspects will enable people to handle Web Services from various perspectives, and help people make full use of more information to locate Web Services. There are several researches focus on organizing Web resources based on their various features, such as Faceted Navigation and RSM [7], but as a kind of special Web resources, Semantic Web Services own some unique features, how to organize them based on their various features in an appropriate way so as to improve the Efficiency of Web Service Discovery is still a problem to be solved.

This paper introduces an ontology-based multi-dimension organization and management model for Semantic Web services space. In this model, semantic Web service set is organized in different feature dimensions, the concepts involved at each dimension come from ontology corresponding to that dimension, and different dimensions can cooperate each other to locate Web Services and reduce the searching space during the process of Web Service discovery, so as to improve the efficiency. This ontology-based multi-dimension organization and management model owns features as follows:

- Reusage of existing knowledge resource.*
- Making full use of more various informations.*
- More highly efficient locating mechanism.*
- Presenting in different granularity or abstract level.*

## 2 Ontology-Based SWS Space Organization and Management Model

### 2.1 Model Overview

Our model acts as a service broker in SOA by building a bridge between applications and Web service entities. It use ontologies to organize Web Services from different perspectives. There are two special mechanisms in our model, one

is the semantic mapping mechanism which can map a Web Service entity to appropriate concepts in ontologies, the other is the Web Services space operation mechanism which can be used to locate and obtain Web Services needed. The architecture of our model is illustrated as Fig. 1 below

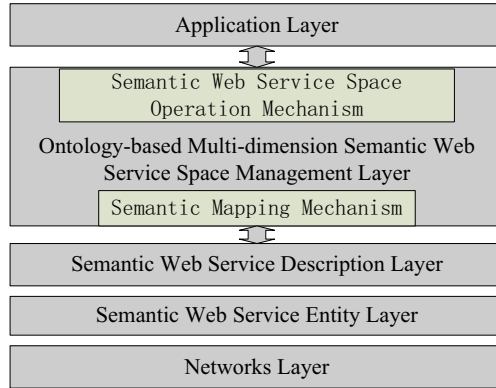


Fig. 1. Architecture of the model

## 2.2 Model Definition

The ontology-based Web service Space organization and Management Model can organize a web service set from several different aspects, for each aspect, an corresponding ontology is maintained and each Web Service entity to be managed by the model is mapped to an appropriate concept in that ontology.

**Definition 1.** *Ontology-based Web Service Space organization and Management model is a multi-dimensional scalable vector space model, which can be denoted as  $M = WSS(N_1, N_2, N_3, \dots, N_n)$ , where*

- *WSS is the name of a Web service space including all the Web services to be managed;*
- *$N_1, N_2, N_3, \dots, N_n$  denote the attribute name of each aspect from which the Web Service set is organized, and they can be regarded as a axis name in an coordinate space separately. To ensure the validation of our model,  $N_1, N_2, N_3, \dots, N_n$  should be mutually orthogonal concepts. For each  $N_i$  ( $1 \leq i \leq n$ ), there is a corresponding ontology  $O_i$  maintained. To a concept set  $\{C_{i1}, C_{i2}, \dots, C_{im}\}$  drawn from  $O_i$ , each concept in it is related to several Web Service entities, if all the concepts constitute a partition of the Web service set in WSS, then they can act as the coordinates of axis  $N_i$  to help locate Web Services, and that can be denoted as  $N_i = \{C_{i1}, C_{i2}, \dots, C_{im}\}$ .*

### 2.3 Model Construction

Based on the definition above, we can obtain an understanding about the model  $M = WSS(N_1, N_2, N_3, \dots, N_n)$  at logical level. Here we introduce the construction of the model. First, through analyzing the features of the descriptions of Web Services and users behaviors, a model designer have to decide that what aspects the Web Services set to be managed should be organized from, that means he should decide the  $N_1, N_2, N_3, \dots, N_n$  in  $WSS$  and he is also demanded to ensure the orthogonality between them. After that, he have to prepare an ontology for each attribute dimension  $N_i$  ( $1 \leq i \leq n$ ). If there is an ontology widely accepted, it can be used in our model directly; if there is no existing ontology to be used directly, an ontology including all the concepts used to describe the Web Services in attribute dimension  $N_i$  has to be constructed by applying expert knowledge in that domain. The mapping relationship between the concepts in an ontology and the Web Services to be managed is established by following two ways: providing a mechanism for the Web Services providers to annotate their Web Services with the ontology existing, or establishing the mapping automatically by analyzing the descriptions of the Web Services to be managed.

### 2.4 Implementation of Model

We use inverted indexing to implement our model, for each ontology, an inverted indexing table is constructed, and the structure of it is as follows:

```

Class ServiceNode {
    String service_id;
    Float relativity;
}
Class InvertedIndex {
    String each concept of  $O_i$ ;
    ArrayList<ServiceNode> service_list;
}
    
```

the relativity in class `ServiceNode` denotes the similarity of the concept used to annotate current Web Service and the concept indexed currently in inverted indexing table, and it can be computed by following formula proposed in [8]:

$$SimCC(C_1, C_2) = \begin{cases} e^{-\alpha l} \cdot \frac{e^{\beta h} - e^{-\beta h}}{e^{\beta h} + e^{-\beta h}} & if(C_1 \neq C_2) \\ 1 & if(C_1 = C_2) \end{cases} \quad (1)$$

Based on the inverted indexing structure, we define the set Operations involved in our model as follows, where  $L_i$  and  $L_j$  are the service lists corresponding to the ontology concept  $O_i$  and  $O_j$  separately:

**Definition 2 (intersection).**  $\forall L_i$  and  $L_j$ , let  $L_i \cap L_j = \{\langle s, r \rangle | s : \langle s, r_i \rangle \in L_i \wedge \langle s, r_j \rangle \in L_j, r : r = \min(r_i, r_j)\}$ .

**Definition 3 (union).**  $\forall L_i$  and  $L_j$ , if  $\langle s, r_i \rangle \notin L_i \wedge \langle s, r_j \rangle \in L_j$ , then  $\langle s, r_j \rangle \in L_i \cup L_j$ ; if  $\langle s, r_i \rangle \in L_i \wedge \langle s, r_j \rangle \notin L_j$ , then  $\langle s, r_j \rangle \in L_i \cup L_j$ ; if  $\langle s, r_i \rangle \in L_i \wedge \langle s, r_j \rangle \in L_j$ , then  $\langle s, \max(r_i, r_j) \rangle \in L_i \cup L_j$ ;

**Definition 4 (cartesian).**  $\forall L_1, L_2, \dots, L_n, \forall \langle s_1, r_1 \rangle \in L_1, \langle s_2, r_2 \rangle \in L_2, \dots, \langle s_n, r_n \rangle \in L_n$ , then  $\langle s_1 + s_2 + \dots + s_n, \frac{1}{n^2} \sum_{i=1}^n r_i \rangle \in L_1 \times L_2 \times \dots \times L_n$ .

### 3 Service Discovery Supported by Model

Here we introduce how to use our model to support Semantic Web Services discovery in the following section.

**Definition 5 (Web Service).** *Web service is a unit that owns some specific attributes and can perform specific functions, which can be defined as a triple  $WS = \langle N, P, D \rangle$ , where*

- $N$  is the ID of a Web service, which is the unique identifier of it.
- $P = \langle p_1, p_2, \dots, p_n \rangle$  denotes the preconditions of a Web service which are the some attributes have to be satisfied before a Web Service to be invoke.
- $D = \langle D_e, D_a \rangle = \langle D_{e1}, \dots, D_{er}, D_{a1}, \dots, D_{as} \rangle$ , where  $D_e$  and  $D_a$  describe the functional properties and non-functional properties of a Web service separately.

**Definition 6 (Web Service Request).** *A Web service request is a property description of a Web service that meets the request of a user, which can be defined as a two-tuple  $WS = \langle Pre, Des \rangle$ , where*

- $Pre$  is the preconditions a user can provide.
- $Des = \langle Des_e, Des_a \rangle = \langle Des_{e1}, \dots, Des_{er}, Des_{a1}, \dots, Des_{as} \rangle$ , where  $Des_e$  and  $Des_a$  describe the functional properties and non-functional properties of a Web service request separately.

**Definition 7.** *The function between any coordinate values  $C_{ij}$  ( $1 \leq j \leq m$ ) on axis  $N_i$  and the Web service set associated with it can be expressed as  $S = W(C_{ij})$ .*

For a Web service request, the properties designated in  $Des$  can be used to locate related Web services quickly in the ontology-based multi-dimension semantic Web service space, and those in  $Pre$  can be used to filter the Web services located.

There are two steps in the atomic level Web services discovery: first, find all Web services that can satisfy the requirements of users, and then check whether all the preconditions of each of them satisfied by the preconditions provided by the user. All Web services that can satisfy the request of user can be acquired by following function:

$$S = \bigcap_{i=1}^r W(Des_{ei}) \cap \bigcap_{j=1}^s W(Des_{aj}) \tag{2}$$

When  $S$  is not empty, it means there are several Web services can satisfy the requirements of users. For each Web services in  $S$ , if the precondition of it can be satisfied, it can be added to the return list; else if none of the Web services in  $S$  can be satisfied, then atomic level service discovery fail.

Our model can also be used to support composition level Web Services discovery, but we won't introduce it in detail here due to the space limitation.

## 4 Conclusion and Future Work

In this paper, an ontology-based multi-dimensional services space organization and management model is proposed. It uses ontology knowledge to organize the semantic Web service from many different perspectives, so that in the process of service discovery, more information can be used fully to limit the semantic web service involved in the operations within a smaller range, and then achieve the purpose of reducing the risk of invalid operations and improving the efficiency of service discovery. At present, this model manages semantic Web service set mainly based on the concepts in ontology, with the emergence of more and more descriptive model, using more effective features to manage Web Service set will be the next research goal.

## References

1. OWL-S: Semantic Markup for Web Services, <http://www.w3.org/Submission/OWL-S/>
2. Roman, D., Keller, U., Lausen, H., de Bruijn, J., Lara, R., Stollberg, M., Polleres, A., Feier, C., Bussler, C., Fensel, D.: Web Service Modeling Ontology. *J. Applied Ontology* 1, 77–106 (2005)
3. METOR-S project site, <http://lstdis.cs.uga.edu/projects/meteor-s/>
4. Li, K., Shui-Guang, D., Ying, L., Jian, W., Hui, W.Z.: Using inverted indexing to facilitate composition-oriented semantic service discovery. *J. Software* 18, 1911–1921 (2007)
5. Richi, N., Bryan, L.: Web Service Discovery with additional Semantics and Clustering. In: *Proceedings of the IEEE/WIC/ACM International Conference on Web Intelligence*, pp. 555–558. IEEE Computer Society, Los Alamitos (2007)
6. Liang, Q., Li, P., Hung, P.C.K., Wu, X.: Clustering Web Services for Automatic Categorization. In: *Proceedings of the 2009 IEEE International Conference on Services Computing*, pp. 380–387. IEEE Computer Society, Los Alamitos (2009)
7. Zhuge, H.: Resource Space Grid: model, method and platform. *J. Concurrency And Computation:Practice And Experience* 16, 1385–1413 (2004)
8. Li, Y.H., Bandar, Z.A., McLean, D.: An approach for measuring semantic similarity between words using multiple information sources. *IEEE Trans. on Knowledge and Data Engineering* 15(4), 871–882 (2003)

# Genetic Algorithm-Based Multi-objective Optimisation for QoS-Aware Web Services Composition

Li Li<sup>1</sup>, Pengyi Yang<sup>2</sup>, Ling Ou<sup>1</sup>, Zili Zhang<sup>1</sup>, and Peng Cheng<sup>1</sup>

<sup>1</sup> Southwest University, Chongqing, 400715, P.R. China

{lily,ouling,zhangz1,chengp}@swu.edu.cn

<sup>2</sup> School of IT, The University of Sydney, Sydney, 2006, Australia

yangpy@it.usyd.edu.au

**Abstract.** Finding an optimal solution for QoS-aware Web service composition with various restrictions on qualities is a multi-objective optimisation problem. A popular multi-objective genetic algorithm, NSGA-II, is studied in order to provide a set of optimal solutions for QoS-based service composition. Experiments with different numbers of abstract and concrete services confirm the expected behaviour of the algorithm.

## 1 Introduction

QoS-aware service composition for meeting non-functional requirements has been widely studied recently [1,2]. However, most existing QoS-aware compositions are simply based on the assumption that multiple criteria, no matter whether they are competing or not, can be handled by the weighted sum approach. Practically, this approach can be very difficult as utility functions or weights are not well known a priori [3].

A motivating scenario may explain the issues raised above clearly. A typical holiday package service includes four service components: *flight Booking*, *hotel Booking*, *car Rental*, and *theme Park Family Pass Booking*. Suppose a large number of candidate Web services with different quality criteria (e.g. *cost*, *response time* and *availability*) are available for every service component, the task of QoS-based service composition is to select the optimal candidate services for each of them. Obviously, qualities such as *cost*, *response time* and *availability* have dimensional constraints (e.g.  $availability \geq 0.6$ ). In addition, there might have some sort of dependent relationships between objectives.

Evolutionary algorithms are very popular approaches in dealing with such kinds of optimisation problems [4]. Genetic algorithms such as the Non-dominated Sorting Genetic Algorithm-II (NSGA-II) [5] and Strength Pareto Evolutionary Approach (SPEA2) [6] are two popular approaches. We will discuss NSGA-II in this paper. The discussion of SPEA2 along with the detailed comparison of these two approaches will be presented in another paper. In this paper, the terms such as criteria and qualities, candidate Web services and concrete Web services are used interchangeably unless otherwise specified.

The rest of the paper is organised as follows. Section 2 introduces the QoS model followed by the problem formalisation. Section 3 discusses the customised NSGA-II. Section 4 adapts NSGA-II to solve QoS-aware service composition with experimental results. Section 5 concludes the paper.

## 2 Problem Description

One of the most promising features of Service Oriented Computing is service composition. The qualities play key role in decisively identifying the best set of services available at runtime. In what follows, we first introduce the QoS model followed by the formalisation of QoS-based composition.

### 2.1 QoS Model

QoS is an integral part of Web services. It is not uncommon that more than one concrete service realising a particular feature is available. Basically, these concrete services are functionally equivalent therefore they can be interchanged. As different concrete services may operate at different QoS measures, these QoS attributes can be used to differentiate a number of functionally equivalent concrete services. In practice, the choice between them is dictated by QoS criteria.

**Table 1.** Aggregation functions

Criteria	Aggregation functions
<i>availability</i>	$\prod_{i=1}^n \prod_{j=1}^m a_{i,j} \cdot x_{j,i}$ , when $x_{j,i} = 1$
<i>reputation</i>	$(\prod_{i=1}^n \prod_{j=1}^m r_{i,j} \cdot x_{j,i})/n$
<i>cost</i>	$\sum_{i=1}^n \sum_{j=1}^m c_{i,j} \cdot x_{j,i}$
<i>reliability</i>	$\prod_{i=1}^n \prod_{j=1}^m e_{i,j} \cdot x_{j,i}$ , when $x_{j,i} = 1$
<i>regulatory</i>	$\min(\sum_{j=1}^m g_{i,j} \cdot x_{j,i})_{i \in \{1 \dots n\}}$
<i>throughput</i>	$(\sum_{i=1}^n \sum_{j=1}^m o_{i,j} \cdot x_{j,i})/n$
<i>latency</i>	$\sum_{i=1}^n \sum_{j=1}^m l_{i,j} \cdot x_{j,i}$
<i>response time</i>	$\sum_{i=1}^n \sum_{j=1}^m t_{i,j} \cdot x_{j,i}$
<i>service capacity</i>	$\min(\sum_{j=1}^m s_{i,j} \cdot x_{j,i})_{i \in \{1 \dots n\}}$
<i>encryption</i>	$(\sum_{i=1}^n \sum_{j=1}^m p_{i,j} \cdot x_{j,i})/n$

Table 1 presents aggregation functions for the computation of QoS qualities used in this paper. Currently 10 quality dimensions are discussed in the travel scenario, however, there is no limitation to the number of characteristics to be handled by the algorithms.

### 2.2 Problem Formulation

Without loss of generality, we assume that all objectives are to be minimised and all equally important. The minimisation<sup>1</sup> of a multi-objective problem with

<sup>1</sup> The maximisation of a multi-objective problem can be implemented as a reverse of minimisation functions.



$k$  objectives is defined as follows ( $n, k \in \mathbb{N}$ ):

to find a vector  $\bar{x} = [x_1, x_2, \dots, x_n]^T$  which *minimise*  $[f_1(\bar{x}), f_2(\bar{x}), \dots, f_k(\bar{x})]^T$  where  $\bar{x} = [x_1, x_2, \dots, x_n]^T$  is a vector of decision variables, which satisfies a series of constraints.

Specifically,  $x_i$  is an integer in this paper, which specifies the index of the matched concrete service. For example, if we have a vector of concrete services in the form of  $\bar{x} = [2, 5, 7, 4]^T$ , it can be represented by a matrix  $\mathbf{x}_{i,j}$  with only the specific elements having values 1s. The definition of the matrix  $\mathbf{x}_{u \times v}$  ( $u, v \in \mathbb{N}$ ) is given below.

$$\mathbf{x}_{i,j} = \begin{cases} 1 & \text{if } j = x_i, \text{ and } x_i \neq 0, \\ 0 & \text{otherwise.} \end{cases}$$

with  $i$  in  $\mathbf{x}_{i,j}$  indicating which abstract service it represents, whilst  $j$  stands for which concrete service has been chosen to match this abstract service. Take the above vector for example, according the definition, we have  $\mathbf{x}_{1,2} = 1$ ,  $\mathbf{x}_{2,5} = 1$ ,  $\mathbf{x}_{3,7} = 1$  and  $\mathbf{x}_{4,4} = 1$ , respectively.

### 3 NSGA-II

In this section, we will concentrate on the core algorithms of NSGA-II. The key algorithms are shown in Figure 1. Our implementation is similar to that developed by Deb et al. [5] with some customisation. The implemented modules are listed in Table 2.

---

MOGA core algorithms

---

```

(01) initiateParameter;
(02) initiateChromosome(popSize);
(03) for  $g = 1$  to generation do
(04)   fastNonDomiSort(chromosome);
(05)   crowdDistanceSort(chromosome);
(06)   nonDominatedSelect(chromosome);
(07)   crossover(chromosome);
(08)   mutate(chromosome);
(09)   for  $i = 1$  to (popSize) do
(10)     for  $j = 1$  to (NumOfObjective) do
(11)       computeFitness(i, j);
(12)     end for
(13)   end for
(14) end for

```

---

**Fig. 1.** NSGA-II Algorithm

**Table 2.** Implemented modules

Modules	Description
<code>initiateParameters()</code>	initiate genetic parameters
<code>initiateChromosome()</code>	generate initiate GA chromosomes randomly
<code>fastNonDominateSort()</code>	non-dominated sorting
<code>crowdDistanceSort()</code>	measure the difference of each solution using the quick sort algorithm
<code>nonDominateSelect()</code>	binary tournament selection by taking account of constraints and solution differences
<code>crossover()</code>	the chromosome crossover algorithm
<code>mutate()</code>	the chromosome mutation algorithm

## 4 Experimental Results

### 4.1 Multi-objectives and Constraints Handling

In order to incorporate multiple constraints, we follow the constraint-awareness selection strategy proposed in [7]. Two types of constraints are considered in the implementation. They are (1) single dimensional constraints; and (2) dependent constraints, respectively.

### 4.2 Parameter Configuration and Experimental Results

The detailed parameters depend on the nature of the problem. With the holiday package example introduced in Section 1, the parameter configuration is shown in Table 3.

**Table 3.** Configuration

Parameters	Values
population size	500
chromosome size	4 and 12, respectively
chromosome encoding	integer encoding scheme
selector	constraint aware tournament selection
crossover	single point (0.7)
mutation	single point (0.03)
termination condition	fixed number of generation
No. of objectives	10
No. of constraints	6
No. of concrete services	30 and 90, respectively

We are interested in finding out the following statistics:

- (Q1) How many non-dominated solutions have been obtained at each generation?
- (Q2) How many solutions are dominated by the immediate succeeding generation for each generation? We expect the result will give us some hints to how the customised NSGA-II is progressing the Pareto front.

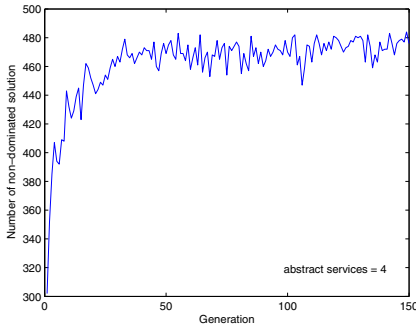
The test data are generated base on some empirical studies. The following are the experiment settings.

- **Setting1:** No. of abstract services = 4, No. of concrete services = 90
- **Setting2:** No. of abstract services = 12, No. of concrete services = 30

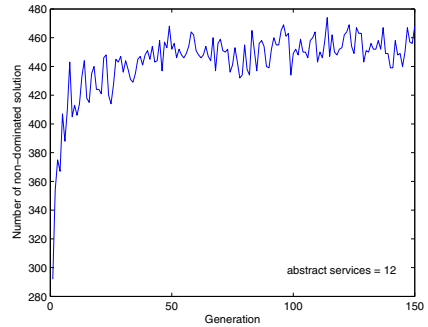
We follow the given parameter configuration in the following two tests.

The first test is focused on answering Q1 by illustrating how many non-dominated solutions are obtained at each generation. We first randomly assign the generated dataset into 4 groups, each with 90 candidate concrete services (i.e., **Setting1** ). We then evaluate the developed algorithms against the data. The outcome is called optimisation results. Then the evaluation work is carried out with another parameter setting, in which the number of the total abstract services is changed to 12, and each abstract service has 30 concrete services (i.e., **Setting2** ).

Figures 2 and 3 illustrate the number of non-dominated solutions obtained at each generation. Apparently, the number of non-dominated solutions increases at nearly the same rate in both figures. These results imply that NSGA-II is able to spread the solutions to the Pareto fronts.



**Fig. 2.** Q1 - setting1



**Fig. 3.** Q1 - setting2

The second test aims at answering Q2 listed above. Experiment settings and the dataset are the same as those in the first test. Figures 4 and 5 demonstrate that at each generation the number of solutions which dominates the solutions of its immediate preceding generation. It is evident that the evolution of the optimisation does not proceed smoothly but fluctuates enormously, which can be seen clearly from the figures. Essentially, Figure 4 and Figure 5 imply that the same pattern exists regardless of the number of abstract services. Moreover, the optimisation process is evidently affected by the increase of the number of abstract services from 4 to 12. The bigger the number of abstract services is, the greater fluctuation the optimisation processes are.

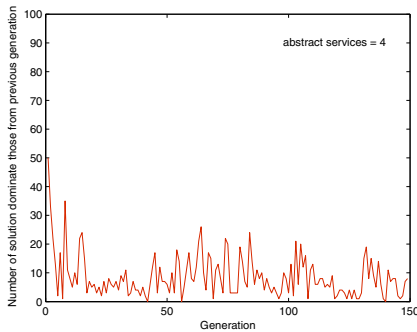


Fig. 4. Q2 - setting1

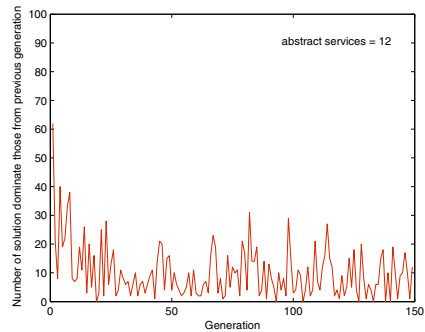


Fig. 5. Q2 - setting2

## 5 Conclusion

We have proposed the use of NSGA-II in order to optimise QoS-based Web service composition. NSGA-II is customised accordingly. The experiment based on the motivating scenario is presented. The experimental results revealed that NSGA-II is a reasonably good fit in solving QoS-aware service composition problems with satisfied convergence and distribution properties.

As future work, we envision comprehensive evaluation of NSGA-II, SPEA2 and particle swarm optimisation with empirical QoS data.

## Acknowledgment

The work is supported in part by the research fund for the Doctoral Program of Southwest University, P. R. China (No. SWU109018, No. SWUB2008006).

## References

1. Ardagna, D., Pernici, B.: Adaptive service composition in flexible processes. *IEEE Trans. on Software Engineering* 33, 369–384 (2007)
2. Hwang, S.Y., Lim, E.P., Lee, C.H., Chen, C.H.: Dynamic web service selection for reliable web service composition. *IEEE Trans. on Services Computing* 1, 104–116 (2008)
3. Fonseca, C.M., Fleming, P.J.: Genetic algorithms for multiobjective optimization: Formulation, discussion and generalization. In: Forrest, S. (ed.) *ICGA 1993*, June 1993, pp. 416–423. Morgan Kaufmann, San Francisco (1993)
4. Zhang, Z., Yang, P., Wu, X., Zhang, C.: An agent-based hybrid system for microarray data analysis. *IEEE IS 24*, 53–63 (2009)
5. Deb, K., Agrawal, S., Pratap, A., Meyarivan, T.: A fast and elitist multiobjective genetic algorithm: NSGA-II. *IEEE Trans. Evolutionary Computation* 6, 182–197 (2002)
6. Zitzler, E., Laumanns, M., Thiele, L.: *Spea2: Improving the strength pareto evolutionary algorithm*. Technical report, ETH, Zürich (2001)
7. Deb, K.: An efficient constraint handling method for genetic algorithms. *Computer Methods in Applied Mechanics and Engineering* 186, 311–338 (2000)

# Knowledge Merging under Multiple Attributes

Bo Wei<sup>1</sup>, Zhi Jin<sup>1,2</sup>, and Didar Zowghi<sup>3</sup>

<sup>1</sup> MADIS, Academy of Mathematics and Systems Science, Chinese Academy of Sciences, Beijing, China

weibo@amss.ac.cn

<sup>2</sup> Key Laboratory of High Confidence Software Technologies, Ministry of Education, Peking University, China

zhijin@amss.ac.cn

<sup>3</sup> Faculty of Engineering and Information Technology (FEIT), University of Technology, Sydney, Australia

idar.zowghi@uts.edu.au

**Abstract.** Knowledge merging is the process of synthesizing multiple knowledge models into a common model. Available methods concentrate on resolving conflicting knowledge. While, we argue that besides the inconsistency, some other attributes may also affect the resulting knowledge model. This paper proposes an approach for knowledge merging under multiple attributes, i.e. *Consistency* and *Relevance*. This approach introduces the discrepancy between two knowledge models and defines different discrepancy functions for each attribute. An integrated distance function is used for assessing the candidate knowledge models.

**Keywords:** Attributes, consistency, knowledge merging, relevance.

## 1 Introduction

Knowledge merging can be essentially considered as a process of belief merging. A typical scenario could be: just like in [1], during the process of patients' diagnosing, each doctor may provide diagnostic information independently. Some of the knowledge conflict mutually that makes patients confused about symptoms' identification. We need a process for obtaining a merging result from different knowledge sources or agents. With the great penetration of Internet, there are now growing number of knowledge sources available online. That makes knowledge merging more important.

Research in knowledge merging can be divided into two main categories. The first is primarily conflict-related. For example, Benferhat *et al.* introduced a computationally effective strategy called *Disjunctive Maxi-Adjustment* (abbr: *DMA*) for conflict resolution among information from multiple sources[10]; Meyer *et al.* proposed the recast techniques for propositional inconsistency management into the description logic setting, and provided algorithms with disjunctive knowledge bases as output[11]. The second is roughly multiple-attribute-related, but ultimately for the purpose of conflicts' resolving. To achieve high accuracy, Brazdil *et al.* designed a method to select some rules from candidate set *TC* and transferred them into integrated theory *TI* (integrated knowledge); Lin *et al.* proposed the integration process of knowledge base with a weight

which demonstrates its relative importance[12]. Delgrande *et al.* proposed a prioritized belief merging method to tackle the information with varying degrees of reliability. [13]. Konieczny *et al.* assumed no preference is cast on different sources and proposed a logical framework for merging information under constraints[14]. Ma defined a set of criteria of interest for knowledge base merging, such as *Traceability*, *Timeliness*, *Lack of redundancy* and *Ease of update*, then assessed them with respect to these criteria[15]. So, consistency is the main focus of most of these work. Our work explicitly raises the multiple attribute-based knowledge merging problem and presents an executable method to get the final result.

We argue that taking conflicts into consideration alone is not sufficient. For the inquirer, it is meaningless to obtain a consistent but not-related knowledge model. This paper focuses on two attributes, i.e. *Consistency* and *Relevance*. We introduce the discrepancy between two knowledge models and define different discrepancy functions for each attribute. Then, an integrated distance function is given as the merging assessment function for assessing the candidate knowledge models. Finally, according to a merging algorithm, the resulting knowledge model will be identified.

The paper is organized as follows: Section 2 introduces attributes in our research background including *Consistency* and *Relevance*. Section 3 presents the main idea of knowledge merging and implementing algorithm. Section 4 gives the conclusion.

## 2 Attribute Selection

The merging process should be defined by a function from the set of all possible collections of consistent knowledge bases to the power set[18]. So, we need all knowledge sets' union, which can be easily identified once knowledge is provided respectively. Each subset of the union can be viewed as a candidate result. Our goal is to use multiple attributes to choose a candidate result as a final one. We call the merging results space of a knowledge set  $S$  containing knowledge from different sources, where its subsets can be viewed as its points. On  $\mathcal{S}$ , we define the distance function to decide the discrepancy between two sets.

**Definition 1.** Let  $S_1, S_2$  be two different points in merging results space and  $S_1 \cap S_2 = \emptyset$ . The distance function  $d(S_1, S_2)$  is defined as:  $d(S_1, S_2) = 1 - \frac{f(S_1 \cap S_2)}{f(S_1 \cup S_2)}$ , where  $f$  satisfies the following properties: (1)  $f(S) \geq 0$ ; (2)  $f(S) = 0$  iff  $S = \emptyset$ ; (3) if  $S_1 \subset S_2$ , then  $f(S_1) \leq f(S_2)$ ; (4) if  $S_1 \cap S_2 = \emptyset$ , then  $f(S_1 \cup S_2) = f(S_1) + f(S_2)$ .

**Proposition 1.** The distance function  $d$  satisfies the following four properties which guarantee  $d$  to be a distance metric: (1)  $d(S_1, S_2) \geq 0$ ; (2)  $d(S_1, S_2) = 0$  iff  $S_1 = S_2$ ; (3)  $d(S_1, S_2) = d(S_2, S_1)$ ; (4)  $d(S_1, S_2) + d(S_1, S_3) = d(S_3, S_2)$ ,  $S_1, S_2, S_3 \in \mathcal{S}$ .

Obviously, the more elements two subsets share, the closer two subsets are. When two subsets completely overlap, the distance becomes 0. The full proof is presented in the appendices.

### 2.1 Consistency Attribute

The merging result should make a good balance of constraints which may contradict or promote mutually. We use "attributes" to demonstrate constraints. Let  $S$  be a knowledge

set containing knowledge from different sources and  $2^S$  be its powerset, an attribute on  $S$  can be viewed as a mapping:  $2^S \rightarrow \mathcal{P}(2^S)$ , where  $\mathcal{P}(2^S)$  is a partition of all possible mapping results so that there exist a unique value that can be assigned to each element of  $2^S$ .

For a given set, its *minimal inconsistent subsets* (abbr.MI) can be identified.  $MI_s$  contain all elements which involves in inconsistencies. We use  $\{MI_1, MI_2, \dots, MI_n\}$  to denote all  $MI_s$  of a set, and  $CORE(S) = \bigcup_{i=1}^n MI_i$ .

**Definition 2.** Let  $S$  be a knowledge set,  $2^S$  be its powerset and  $S' \subseteq 2^S$ . The Consistency  $CT(S')$  of  $S'$  is defined as:  $CT(S') = \frac{1}{n} \sum_{i=1}^n d(S', MI_i(S))$ .

It is noticeable that cardinality function  $(\cdot)$  satisfies five properties of  $f$  in  $d$  of Def. 3. So, if  $f$  in  $d$  is asserted by  $\cdot$  of  $S$ ,  $CT(S') = 1 - \frac{1}{n} \sum_{i=1}^n \frac{|S' \cap MI_i(S)|}{|S \cup MI_i(S)|}$ . Furthermore, the *Consistency* of a set decreases when the set gets close to  $MI_i$ . Note that *Consistency* here makes a little difference from what have been discussed in the traditional literature[12,11,2,3]. We focus on exploring potential inconsistencies. So the concept of *Consistency* not only captures explicit inconsistencies, but latent inconsistencies as well.

To merge knowledge under multiple attributes, we need the optimal result from one single attributes' side.  $CT(S)$  and  $CT(MI_i)$  ( $i=1, \dots, n$ ) are viewed as the best/worst merging results upon *Consistency* attribute respectively.

### 2.2 Relevance Attribute

For relevance attribute, we use the “*Relevance*” to capture the degree to which the resulting knowledge set answers the request. Note that, the *Relevance* attribute discussed in this paper is semantic-based. The closer the knowledge obtained is to the requesters' meaning, the more relevant the knowledge obtained is. Hence we adopt *semantic similarity* evaluation to obtain final relevance assignment for each piece of knowledge. We apply *Latent Semantic Analysis* (abbr. LSA) into our knowledge merging process. The reasons are twofold. Firstly, LSA is one of the two semantic similarity evaluation methods, whose performances are claimed to be mostly closed to human judgement [19,20]. Secondly, LSA supports the evaluation between one term(keyword) and many documents. Thus it suits our goal in finding the most relevant result whereby in many cases, inquirers provide only one keyword in order to get feedback from knowledge sources[21]. *Relevance* of knowledge set is always reflected by its individual pieces of knowledge. If each piece of knowledge is highly relevant to the request, the whole set can be said to be highly relevant. To evaluate the complete relevance of a knowledge set, we introduce a concept of *total relevance*.

**Definition 3.** Let  $S = A_1 \cup A_2 \cup \dots \cup A_n$  be a knowledge set.  $TR(S)$  is called the total relevance of  $S$  to the request, if  $TR(S) = \sum_{i=1}^n r_i$  where  $r_i \in (0, 1]$  is a semantic-similarity score of  $A_i$  given by LSA.

Obviously, the complete set  $S$  has the highest relevance, which can serves as a baseline to evaluate other subsets. It is noticeable that total relevance  $TR$  satisfies five properties of  $f$  in  $d$  of Def. 3.

**Definition 4.** Let  $S$  be a knowledge set,  $2^S$  be its powerset and  $S' \in 2^S$ . The Relevance  $RL(S')$  of  $S'$  to the request is defined as:  $RL(S') = \frac{1}{d(S', S)}$ .

If  $f$  in  $d$  is asserted by TR,  $RL(S') = \frac{TR(S', S)}{TR(S' \cup S)}$ . RL increases as the cardinality of a set increases. Similarly,  $RL(S) = \frac{1}{d(S, \emptyset)}$  and  $RL(\emptyset) = 1$  are viewed as the best/worst merging results upon Relevance attribute respectively.

### 3 Knowledge Merging Process

Different attributes generally have different priorities in decision making. Their priorities can be represented by weights. So, it is reasonable to assess each candidate result by calculating the distances between the candidate result and the best/worst result based on the mutual weights of different attributes. The candidate result far from the worst and close to the best will be chosen as the final result.

**Definition 5.** Let  $S$  be a knowledge set,  $2^S$  be its powerset and  $S' \in 2^S$ .  $CT, RL$  are weights of Consistency and Relevance respectively, and  $d(S', S) = \frac{CT \cdot d(S', S) + RL \cdot d(S', S)}$  is called the distance of the  $S'$  to the positive optimal result.

**Definition 6.** Let  $S$  be a knowledge set,  $2^S$  be its powerset and  $S' \in 2^S$ .  $CT, RL$  are weights of Consistency and Relevance respectively, and  $d(S', S) = \frac{CT \cdot \frac{1}{n} \sum_{i=1}^n d(S', S_i) + RL \cdot d(S', S)}$  is called the distance of the  $S'$  to the negative optimal result.

Using  $d_+$  and  $d_-$ , the merging assessment function can be defined:

**Definition 7.** Let  $S$  be a knowledge set,  $2^S$  be its powerset and  $S' \in 2^S$ .  $d_+$  and  $d_-$  are distances of the  $S'$  to the positive optimal and negative optimal results respectively.  $f(S')$  is called the merging assessment function if  $f(S') = \frac{d_-(S')}{d_+(S') + d_-(S')}$ .

Obviously, the greater  $f(S')$  value means the higher degree of relatedness to the positive optimal result.

```

Input: a knowledge set  $S = \{A_1, \dots, A_n\}$  and its relevance vector  $v_R(S) = \{r_1, \dots, r_n\}$ 
Output:  $(survival \cup positive)(S) = (survival \cup positive)$ 
for  $i, j \in \mathbb{N}$  do
     $survival = \emptyset, positive = S \text{ CORE}(S);$ 
     $A_{ij} = \{M_i, r_{ij}, \max_k r_{ik}\};$ 
     $(survival \cup A_{ij}) = \max\{(survival \cup A_{ij})\};$ 
    if  $n \in \mathbb{N}, M_n \notin survival \cup A_{ij};$ 
         $(survival \cup A_{ij}) = (survival) \text{ then}$ 
             $survival = survival \cup A_{ij};$ 
             $M_i = M_i \cup \{A_{ij}\};$ 
        end
    end
end
return  $(survival \cup positive)(S) = (survival \cup positive)$ 

```

**Algorithm 1.** Knowledge merging algorithm based on Consistency and Relevance



Knowledge merging assessment function  $(S')$  can guide us to choose the optimal knowledge merging result, but it presents us with a very high computational complexity for  $powset(S) = 2^S$ . It is not practical to directly apply  $(S')$  to evaluate every candidate set. According to the definition, adding knowledge from  $S \in CORE(S)$  can increase *Relevance* and *Consistency* of the set at the same time. So, choosing knowledge from  $CORE(S)$  is the key issue to consider. Hence if  $S$  is substituted by  $CORE(S)$ , then  $(S')$   $\frac{CT_{CT(S')} \cdot RL_{RL}}{CT_{CT(S')} \cdot RL_{RL} + 1}$ . Using the formula above we can identify which knowledge should be selected from  $CORE(S)$ . Algorithm 1 implements the whole knowledge merging process.

### 4 Conclusion

Knowledge merging is a prevailing topic which has attracted much attention for decades. This paper focuses on a specific form of merging process where multiple attributes of consistency and relevance are considered in parallel. The overall goal is to present a merging result that is as consistent and as relevant to the inquirer’s context as possible. So, we have used attributes of *Consistency* and *Relevance* to conduct a systematic evaluation of each candidate result. These attributes are unified on to the distance metric between two sets. Naturally, distances from candidate results to optimal result and negative optimal result are employed to identify whether the candidate result is close enough to the optimal one. We give an algorithm to implement our idea.

**Acknowledgements.** This work is supported by the National Natural Science Funds for Distinguished Young Scholar under Grant No.60625204; the National Grand Fundamental Research Program of China under Grant No. 2009CB320701, the Key Projects of National Natural Science Foundation of China under Grant Nos. 90818026, and the International Science Linkage Research Grant under the Australia-China Special fund for Science and technology.

### References

1. Lin, J., Mendelzon, A.O.: Merging databases under constraints. *Int. J. of Cooperative Information Systems* 7(1), 55–76 (1998)
2. Knight, K.: Measuring inconsistency. *Journal of Philosophical Logic* 31(1), 77–98 (2002)
3. Grant, J., Hunter, A.: Measuring inconsistency in knowledgebases. *Journal of Intelligent Information Systems* 27, 159–184 (2006)
4. Hunter, A., Konieczny, S.: Approaches to measuring inconsistent information. In: Bertossi, L., Hunter, A., Schaub, T. (eds.) *Inconsistency Tolerance*. LNCS, vol. 3300, pp. 191–236. Springer, Heidelberg (2005)
5. Mu, K., Jin, Z., Lu, R., Liu, W.: Measuring inconsistency in requirements specifications. In: Godo, L. (ed.) *ECSQARU 2005*. LNCS (LNAI), vol. 3571, pp. 440–451. Springer, Heidelberg (2005)
6. Hunter, A., Konieczny, S.: Measuring inconsistency through minimal inconsistent sets. In: *Principles of knowledge representation and reasoning: Proceedings of the eleventh international conference (KR 2008)*, pp. 358–366 (2008)

7. Grant, J., Hunter, A.: Analysing inconsistent first-order knowledge bases. *Artificial Intelligence* 172, 1064–1093 (2008)
8. Hunter, A., Konieczny, S.: Approaches to measuring inconsistent information. In: Bertossi, L., Hunter, A., Schaub, T. (eds.) *Inconsistency Tolerance*. LNCS, vol. 3300, pp. 189–234. Springer, Heidelberg (2005)
9. Reiter, R.: A theory of diagnosis from first principles. *Artificial Intelligence* 32, 57–95 (1987)
10. Benferhat, S., Kaci, S., Le Berre, D., Williams, M.A.: Weakening conflicting information for iterated revision and knowledge integration. In: *IJCAI*, pp. 109–118 (2001)
11. Meyer, T., Lee, K., Booth, R.: Knowledge integration for description logics. In: *AAAI*, pp. 645–650 (2005)
12. Lin, J.: Integration of weighted knowledge bases. *Artif. Intell.* 83(2), 363–378 (1996)
13. Delgrande, J., Dubois, D., Lang, J.: Iterated revision as prioritized merging. In: *KR 2006*, pp. 210–220 (2006)
14. Konieczny, S., Pino Perez, R.: Merging information under constraints: a logical framework. *Journal of Logic and Computation* 12(5), 773–808 (2002)
15. Ma, Z.M.: Propositional knowledge bases merging. In: *I-KNOW* (2005)
16. Gregoire, E., Konieczny, S.: Logic-based approaches to information fusion. *Information Fusion* 7, 4–18 (2006)
17. Hunter, A.: Merging potentially inconsistent items of structured text. *Data and Knowledge Engineering* 34, 305–332 (2000)
18. Gabbay, D.M., Pigozzi, G., Rodrigues, O.: Belief revision, belief merging and voting. In: *The Seventh Conference on Logic and the Foundations of Games and Decision Theory*, University of Liverpool, pp. 71–78 (2006)
19. Gabrilovich, E., Markovitch, S.: Computing semantic relatedness using wikipedia-based explicit semantic analysis. In: *Proceedings of The 20th International Joint Conference on Artificial Intelligence (IJCAI)*, Hyderabad, India (January 2007)
20. Milne, D.: Computing semantic relatedness using wikipedia link structure. In: *NZ CSRSC* (2007)
21. Landauer, T.K., Foltz, P.W., Laham, D.: Introduction to latent semantic analysis. *Discourse Processes* 25, 259–284 (1998)

## Appendix

*Proof.*  $d$  is a distance metric.

(1) If  $(S_1 \cup S_2) \subset (S_1 \cup S_2)$ , then we have  $f(S_1 \cup S_2) = f(S_1 \cup S_2)$ ,  $d = 0$ ; If  $(S_1 \cup S_2) \not\subset (S_1 \cup S_2)$ , then we have  $f(S_1 \cup S_2) < f(S_1 \cup S_2)$ ,  $d > 0$ . (2) If  $S_1 \cup S_2$ , then  $S_1 \cup S_2 \subset S_1 \cup S_2$ ,  $d(S_1 \cup S_2) = 0$ . If  $d(S_1 \cup S_2) > 0$ , then  $f(S_1 \cup S_2) < f(S_1 \cup S_2)$ . If  $(S_1 \cup S_2) \subset (S_1 \cup S_2)$ , then  $f(S_1 \cup S_2) = f(S_1 \cup S_2)$ , which should not hold. So  $S_1 \cup S_2 \subset S_1 \cup S_2$ . Consequently,  $S_1 \cup S_2 \subset S_1 \cup S_2$ . (3)  $d(S_1 \cup S_2) = d(S_2 \cup S_1)$  is obvious. (4) Transform the formula  $d(S_1 \cup S_2) = d(S_1 \cup S_3) = d(S_3 \cup S_2)$ , we get the inequality:  $\frac{f(S_3 \cup S_2)}{f(S_3 \cup S_2)} + \frac{f(S_1 \cup S_3)}{f(S_1 \cup S_3)}$

$1 - \frac{f(S_1 \cup S_2)}{f(S_1 \cup S_2)}$ .  $S_1, S_2$ , and  $S_3$  are assumed to be divided into seven non-intersecting parts  $v_1, \dots, v_7$ . Use notation  $f_i$  to denote  $f(v_i)$  for explanatory convenience. So, the inequality mentioned above could be simplified as  $\frac{f_5 f_7}{f_1 f_4 f_5 f_6 f_7 f_3} + \frac{f_6 f_7}{f_2 f_4 f_5 f_6 f_7 f_3} - 1$   
 $\frac{f_4 f_7}{f_1 f_2 f_4 f_5 f_6 f_7}$ . For  $\frac{f_5 f_7}{f_1 f_4 f_5 f_6 f_7 f_3} - \frac{f_6 f_7}{f_2 f_4 f_5 f_6 f_7 f_3} + \frac{f_5 f_7}{f_1 f_4 f_5 f_6 f_7} - \frac{f_6 f_7}{f_2 f_4 f_5 f_6 f_7}$ .  
 So we need to prove  $\frac{f_5 f_7}{f_1 f_4 f_5 f_6 f_7} - \frac{f_6 f_7}{f_2 f_4 f_5 f_6 f_7} + 1 - \frac{f_4 f_7}{f_1 f_2 f_4 f_5 f_6 f_7}$ . Transforming that inequality, we can get new inequality:  $\frac{f_5 f_7}{f_1 f_4 f_5 f_6 f_7} * f_2 - \frac{f_6 f_7}{f_2 f_4 f_5 f_6 f_7} * f_1 - f_1 - f_2 - 2f_4$ . It holds obviously.

# Feature Selection Based on Mutual Information and Its Application in Hyperspectral Image Classification

Na Yao<sup>1,2</sup>, Zongjian Lin<sup>2</sup>, and Jingxiong Zhang<sup>1</sup>

<sup>1</sup> School of Remote Sensing Information Engineering, Wuhan University,  
129 Luoyu Road, Wuhan, China

<sup>2</sup> Chinese Academy of Surveying and Mapping, 16 Beitaiping Road, Beijing, China  
joviayoyo@gmail.com, lincasm@casm.ac.cn, jxzhang@whu.edu.cn

**Abstract.** This paper investigates mutual information-based feature selection for high dimensional hyperspectral imagery, which accounts for both the relevance of features on classes and the redundancy among features. A representative method shortly known as min-redundancy and max-relevance (mRMR) was adopted and compared with a baseline method called Max-Relevance (MR) in experiments with AVIRIS hyperspectral data. Supervised classifications were also carried out to identify classification accuracies obtainable with hyperspectral data of reduced dimensionality through five different classifiers. The results confirm that mRMR is more discrimination-informative than MR in feature selection due to the additional redundancy analysis. Different classifiers with different accuracies manifest that a more impact but more informative subset may exist. However, the intrinsic dimensionality which indicates the optimal performance of a classifier remains an issue for further investigation.

**Keywords:** Hyperspectral image, feature selection, mutual information, relevance, redundancy, supervised classification.

## 1 Introduction

With the advent of hyperspectral remote sensors with hundreds of contiguous narrow spectral bands in the last two decades, it is reasonable to expect the increased spectral bands to contain more information and hence to be able to detect more classes with more accuracy [1]. However, a number of methods suitable for multispectral data cannot be simply transplanted to hyperspectral data [1-3].

Consider the case of supervised classification. Although increased spectral bands potentially provide more information about class separability, this positive effect is diluted by poor parameter estimation of supervised classification [2]. This is due to what is known as the Hughes phenomenon, i.e., added features may actually degrade the performance of a classifier if the number of training samples used to design the classifier is small relative to the number of features [2-5].

The number of training samples required by supervised classification is a function of feature dimension, which is termed as “curse of dimensionality” [4]. Studies of high dimensional feature space suggest that the volume of a hypercube concentrates in the

corners, and the volume of a hypersphere concentrates in an outside shell [1-3]. Therefore, we can reduce dimensionality without incurring significant loss of information and separability among classes [3].

Feature selection is a means of dimensionality reduction. Given a set  $X$  of  $n$  features, the problem of feature selection is to select a subset  $Y$  of size  $m$  from  $X$  that leads to the smallest classification error [4]. Feature selection algorithms can be classified into two main groups: filters and wrappers. Due to the computational efficiency, the filter methods are very popular for high dimensional data [6].

A main issue of dimensionality reduction concerns criterion functions [4], which measure how good a specific subset can be in discriminating between classes [7]. Several different criteria have been used for evaluating the goodness of a feature set, including distance measures, dependency measures, consistency measures, information measures and classification error measures [6]. Although information measures have been comprehensively studied [6-12], they have seldom been applied in remote sensing.

The rest of the paper is organized as follows. Section 2 presents the two existing feature selection methods based on mutual information. Section 3 presents experimental results with hyperspectral data. Section 4 and Section 5 are discussions and conclusions, respectively.

## 2 Feature Selection Based on Mutual Information

In feature selection, the relevant features have important information regarding the output, whereas the irrelevant features contain little information regarding the output [10]. Mutual information provides a feasible way to measure the relevance of two or more variables.

Given a discrete random variable  $C$ , representing the class labels, the initial average uncertainty of  $C$  can be quantified by entropy  $H(C)$  as:

$$H(C) = - \sum_{c \in C} p(c) \log p(c), \quad (1)$$

where  $p(c)$  is the probability distribution for variable  $C$ . As for the hyperspectral data in this paper, however, the feature vector exemplified by radiance response is continuous. Thus, after knowing the feature vector  $\mathbf{f}$ , the remaining uncertainty is termed as conditional entropy:

$$H(C|F) = - \int p(\mathbf{f}) \left[ \sum_{c \in C} p(c|\mathbf{f}) \log p(c|\mathbf{f}) \right] d\mathbf{f}, \quad (2)$$

where  $p(c|\mathbf{f})$  is the conditional probability for class  $c$  given the input vector  $\mathbf{f}$ .

The amount by which  $H(C)$  is decreased by  $H(C|F)$  is, by definition, the mutual information  $I(C;F)$ , which is therefore the amount by which the knowledge provided by the feature vector decreases the uncertainty about the class [9]. Taking the symmetric characteristic of mutual information into consideration, we obtain

$$I(C; F) = I(F; C) = H(C) - H(C|F) = \sum_{c \in C} \int p(c, \mathbf{f}) \log \frac{p(c, \mathbf{f})}{p(c)p(\mathbf{f})} d\mathbf{f}, \quad (3)$$

where  $p(c, \mathbf{f})$  is the joint probability density and  $p(c)$  together with  $p(\mathbf{f})$  is the marginal probability density.

A natural idea for feature selection is to find a feature set  $S$  with features  $\{f_i, i=1, 2, \dots, m\}$ , i.e., a feature vector  $\mathbf{f}$ , which jointly have the largest dependency on the target class  $c$ . In other words, the purpose of feature selection is to find a feature vector  $\mathbf{f}$  that maximize  $I(\mathbf{f}; c)$ .

However, it is practically impossible to obtain the theoretical value of  $I(\mathbf{f}; c)$  due to the difficulty in accurately estimating multivariate probability density functions (pdfs) and calculating integration. Instead, a heuristic method only computing  $I(f; c)$  and  $I(f_i; f_j)$  is preferred and widely adopted, where  $f_i$  and  $f_j$  are individual features.

Thus, a scheme called Max-Relevance (MR) is proposed [8-12]. Given the initial feature set  $F$  with all features, the MR criterion is to search features satisfying Eq.4:

$$\max J(S, c), \quad J = \frac{1}{|S|} \sum_{f_i \in S} I(f_i, c), \quad (4)$$

where  $|S|$  is the cardinality of the selected feature subset  $S$  and  $J(\cdot)$  is an information measure that quantifies the relevance. According to this criterion, a new feature  $f_i$  is selected with maximal relevance on the target class  $c$ .

However, a candidate feature from the unselected set  $(F-S)$  with maximum relevance with class  $c$  may be redundant with those already selected features in set  $S$  [13]. Therefore, the following Min-Redundancy condition should be added [12]:

$$\min K(S, c), \quad K = \frac{1}{|S|^2} \sum_{f_i \in S} I(f_i; f_j), \quad (5)$$

where  $K(\cdot)$  is a criterion that quantifies the relevance.

The criterion combining the aforementioned relevance analysis and redundancy analysis is called min-redundancy and max-relevance (mRMR) [12]. Its purpose is to optimize the following condition:

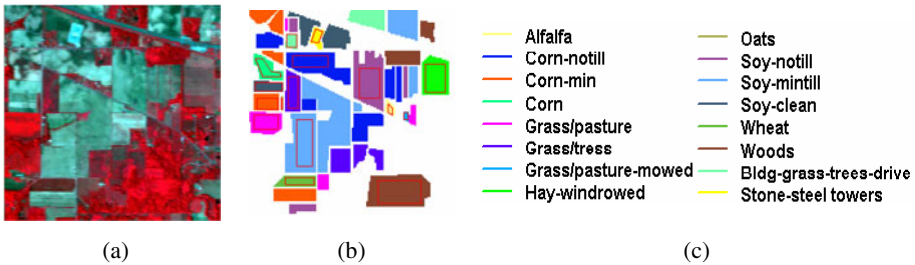
$$f_j \in F - S_{m-1} \left[ I(f_j; c) - \frac{1}{|S_{m-1}|} \sum_{f_i \in S_{m-1}} I(f_j; f_i) \right]. \quad (6)$$

### 3 Data Description and Experiments

The hyperspectral data used in this experiment was acquired in June 1992 by AVIRIS, which covers a 100 square mile area in Northwestern Indiana over the Indian Pine Test site. The data set is made up of 145 by 145 pixels with 220 spectral bands (Fig.1(a)). Meanwhile, a detailed ground truth map featuring 16 land cover classes is available for

sampling (Fig.1(b)). In our experiment, 983 training samples and 1964 testing samples (nearly twice the size of training samples) are selected for further classification accuracy assessment. Fig.1(c) shows the 16 land cover classes.

In order to avoid the integral calculations for continuous radiance response vector, a simple discretization method using standard deviation  $\sigma$  and mean  $\mu$  was applied to the feature set. Furthermore, five different classifiers (i.e., Naive Bayes (NB) classifier, Generalized Linear Model (GLM), Regression Tree (RT), Support Vector Machine (SVM), Linear Discriminant Analysis (LDA)) are chosen to test the adaptability and effectiveness when both mRMR criterion and MR criterion are utilized to hyperspectral data.



**Fig. 1.** Data: (a) an AVIRIS hyperspectral image; (b) a corresponding ground truth map of (a) with the red rectangles representing sampling areas; (c) 16 land cover classes

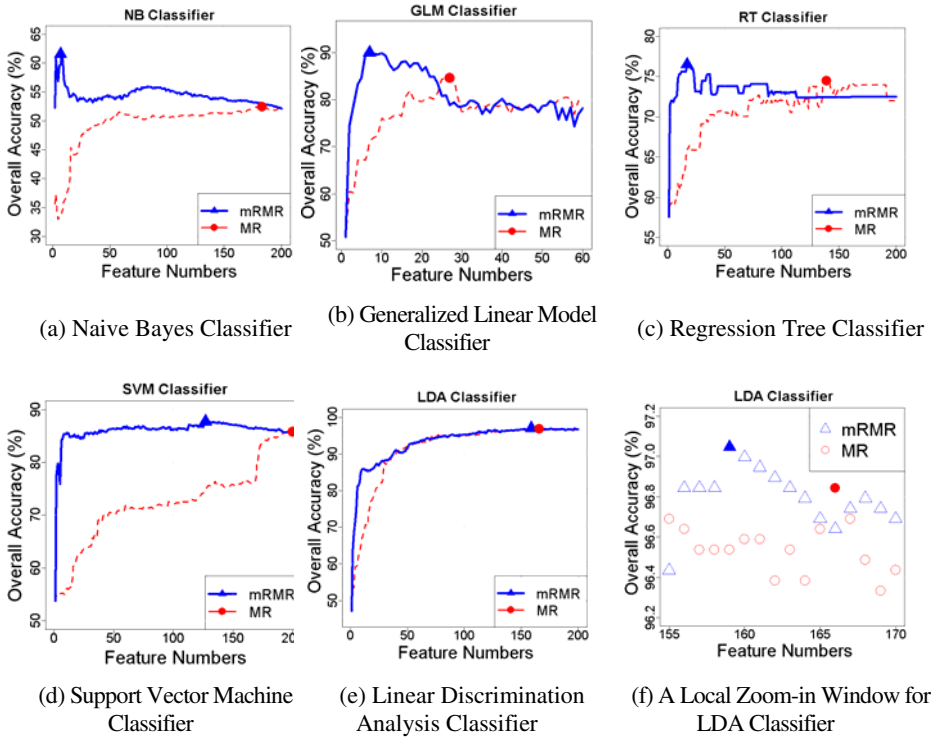
## 4 Discussions

Fig.2, (a)~(e) show the classification results of the aforementioned five classifiers. The ordinates represent overall accuracy and the abscissas indicate the sequential features selected by the mRMR criterion and MR criterion. The thick solid blue lines and thin dashed red lines demonstrate the overall accuracies based on the features selected by the mRMR criterion and MR criterion, respectively. For the LDA classifier, a zoom-in window in Fig.2(f) is provided for better discerning of the two lines in Fig.2(e). For all the five classifiers, the blue solid triangles label the maximal overall accuracies corresponding to the mRMR criterion, while the red solid circles label those corresponding to the MR criterion.

In general, the mRMR obtains higher accuracy with less features than the MR criterion, as the latter focuses on extracting the most “informative” features without considering the redundancy between features. In other words, features selected later may fail to complement previously selected feature by the MR criterion. This is confirmed by the classification results of the five different classifiers.

Classification performance depends on four factors, i.e., class separability, training sample size, dimensionality, and classification algorithms [2]. It was shown in Fig.2 that the performances of different classifiers are various given the same training sample size. One reason is that available information may be sufficient to resolve all ambiguities but a certain classifier “wastes” some of it. And for each classifier, the

differences in accuracies tend to become smaller as more and more features are added due to the fact that most of the essential bands are already included. Indeed, the training sample sizes will rapidly become the prevailing factor which restricts the classification accuracy as the dimensionality increases.



**Fig. 2.** Classification accuracies of five different classifiers

In Fig.2, the LDA classifier performs best mainly due to the fact that it involves a dimension reduction procedure. This implies that a more compact feature set may be further selected. For example, a two-stage procedure which integrates a wrapper to select a reduced subset of features is widely used [12], indicating that a smaller but more informative subspace is embed in the feature set selected by the two criteria. Due to the different performances of various classifiers, however, it is doubtful that feature subsets of reduced dimensionality will be easily determined. As for the optimal dimensionality, i.e., the intrinsic dimensionality [4], open issues remain for further investigation. One of the issues concerns class separability, which reflects the nature of a data set and seems to limit the optimum accuracy achievable with a classifier [2].

## 5 Conclusions and Future Work

The experimental results have shown that the feature selection based on mutual information is a promising dimensionality reduction approach for hyperspectral data.

Further study should be directed towards directly estimating the probability density for continuous data set, extracting more compact subsets from features sequentially selected by reckoning that different classifiers with different accuracy assessment will probably result in conflicting indices, and determining the optimal dimensionality of the subset through knowledge of the underlying class separability.

## Acknowledgments

The authors would like to thank Prof. D. A. Landgrebe for the accessible resources of hyperspectral data, and Dr. Hanchuan Peng for sharing his research work in this field. The research was partially funded by the National Basic Research Program of China (No. 2007CB714402 5).

## References

1. Landgrebe, D.: Some Fundamentals and Methods for Hyperspectral Image Data Analysis. In: SPIE International Symposium on Biomedical Optics (Photonics West), San Jose California (1999)
2. Hsieh, P.F., Landgrebe, D.: Classification of High Dimensional Data. PhD Thesis and School of Electrical & Computer Engineering Technical Report TR-ECE 98-4 (1998)
3. Luis, J., Landgrebe, D.: Hyperspectral Data Analysis and Feature Reduction Via Projection Pursuit. *IEEE Transactions on Geoscience and Remote Sensing* 37(6), 2653–2667 (1999)
4. Jain, A.K., Duin, R.P.W., Mao, J.C.: Statistical Pattern Recognition: a Review. *IEEE Transaction on Pattern Analysis and Machine Intelligence* 22(1), 4–37 (2000)
5. Landgrebe, D.: Multispectral Data Analysis: A signal Theory Perspective (2005), [http://cobweb.ecn.purdue.edu/~biehl/MultiSpec/Signal\\_Theory.pdf](http://cobweb.ecn.purdue.edu/~biehl/MultiSpec/Signal_Theory.pdf)
6. Estévez, P.A., Tesmer, M., et al.: Normalized Mutual Information Feature Selection. *IEEE Transaction on Neural Networks* 20(2), 189–200 (2009)
7. Ali, E.A., Abdeljalil, E.O., Driss, A.: A Powerful Feature Selection Approach Based on Mutual Information. *IJCSNS International Journal of Computer Science and Network Security* 8(4), 116–121 (2007)
8. Battiti, R.: Using Mutual Information for Selecting Features in Supervised Neural Net Learning. *IEEE Transactions on Neural Networks* 5(4), 537–551 (1994)
9. Li, Y.F., Xie, M., Goh, T.N.: A Study of Mutual Information Based Feature Selection for Case Based Reasoning in Software Cost Estimation. *Expert Systems with Applications* (2008)
10. Oveisi, F., Erfanian, A.: A Minimax Mutual Information Scheme for Supervised Feature Extraction and Its Application to EEG-Based Brain-Computer Interfacing. *EURASIP Journal on Advances in Signal Processing* 2008, 1–8 (2008)
11. Kwak, N., Choi, C.H.: Input Feature Selection for Classification Problems. *IEEE Transactions on Neural Networks* 13(1), 143–160 (2002)
12. Peng, H.C., Long, F.H., Ding, C.: Feature Selection Based on Mutual Information: Criteria of Max-Dependency, Max-Relevance, and Min-Redundancy. *IEEE Transaction on Pattern Analysis and Machine Intelligence* 27(8), 1226–1238 (2005)
13. Colannino, J., Juban, J.: The Best K Measurements Are Not the K Best (2004), <http://cgm.cs.mcgill.ca/~athens/cs644/Projects/2004/JustinColannino-JeremieJuban/>



# Static, Dynamic and Semantic Dimensions: Towards a Multidisciplinary Approach of Social Networks Analysis

Christophe Thovex and Francky Trichet

LINA, University of Nantes

Laboratoire d'Informatique de Nantes Atlantique (UMR-CNRS 6241)

2 rue de la Houssiniere, BP 92208 - 44322 Nantes cedex 03, France

{christophe.thovex, franky.trichet}@univ-nantes.fr

<http://www.lina.univ-nantes.fr/-COD-.html>

**Abstract.** The objective of our work is to extend static and dynamic models of Social Networks Analysis (SNA), by taking conceptual aspects of enterprises and institutions social graph into account. The originality of our multidisciplinary work is to introduce abstract notions of electro-physic to define new measures in SNA, for new decision-making functions dedicated to Human Resource Management (HRM). This paper introduces a multidimensional system and new measures: (1) a *tension* measure for social network analysis, (2) an electrodynamic, predictive and semantic system for recommendations on social graphs evolutions and (3) a *reactance* measure used to evaluate the individual *stress* at work of the members of a social network.

**Keywords:** Social networks, static analysis, dynamic analysis, semantic analysis, ontology.

## 1 Introduction

Social Networks Analysis - SNA - is now extending to enterprises so as to provide new management tools devoted to work organisation, workforce and human resources management. A social network is usually formalised with a (not)oriented, (not)labelled and weighted graph. From such a structure, we differentiate three kinds of SNA: *static SNA*, *dynamic SNA* and *semantic SNA*. *Static SNA* studies the state  $S$  of a social graph at a time  $t$  through models and measures mainly dedicated to structures [4]. *Dynamic SNA* studies the evolution of a social graph from a state  $S$  at a time  $t$ , to a state  $S'$  at a time  $t'$ . Depending on well-known topologies such as so-called random graph [3] or scale-free graph [1], a social network owns a characterised behaviour. *Semantic SNA* studies the conceptual aspects of a social graph, based on the principles underlying conceptual graphs and semantic networks theory [8]. Semantic SNA refers to Semantic Web, Ontology Engineering and logical inferences, in correlation with cognitive sciences.

The main objective of our work is to exhibit multidimensional synergies between the static, dynamic and semantic aspects in Enterprises and Institutions

Social Networks - EISN. The specificities of EISN are: (1) social graphs composed of up to 100 000 nodes, (2) endogenous data restricted to a few specific and connate domains and (3) intensive collaborative work with trade oriented information sharing. Our work is funded by the French State Secretariat for prospective and development of the digital economy, in the context of the SOCIOPRISE project. It is developed in collaboration with a French IT service and software engineering company, *OpenPortal Software*, which provides industry-leading software dedicated to human capital management.

The rest of this paper is structured as follows. Section 2 introduces the principles and methods used for *static SNA*, *dynamic SNA* and *semantic SNA*. Section 3 presents the approach we advocate to integrate static, dynamic and semantic SNA. Our contributions are based on (1) a bridge-building between radio-electronic principles to complete static analysis or dynamic analysis, and (2) a bridge-building between our new physical measures of SNA and knowledge engineering. Our work is devoted to Enterprises and Institutions Social Networks Analysis - EISNA - and should be applied to improve performances while reducing psychosocial troubles<sup>1</sup>.

## 2 Unidimensional Approaches

### 2.1 Static Analysis

The centrality measures are based on the comparison of a vertex degree or proximity to those of the graphs, neighbours or distant ones - *e.g. centrality of power, centrality of prestige, centrality of closeness*. *Betweenness* of a vertex defines how an individual is important to interconnect other members of the social graph (non-oriented). Based on [4], we formalise it as follows:

$$\forall i \neq u \neq j, \sigma(i, u, j) > 0, I_u = \sum_{(i,j)} \frac{\sigma(i, u, j)}{\sigma(i, j)} \quad (1)$$

where  $\sigma(i, j)$  is the count of shortest chains between  $i$  and  $j$ ,  $\sigma(i, u, j)$  is the count of shortest chains between the vertices  $i$  et  $j$  crossing  $u$ . The ratio  $\sigma(i, u, j)$  by  $\sigma(i, j)$  is cumulated for the  $(i, j)$  where  $\sigma(i, u, j) > 0$ .

### 2.2 Radio-Electrical Principles for Dynamic EISNA

According to an intuitive observation of digital social networks behaviours, we allege that information flows and nodes in a social network are, in a pure semantic and abstract view, close to an electro-physic system and to radio-electrical principles. We assimilate the edges of a social graph to conductors transporting electrical flows and assume that percentages of numerical communication marks between individuals (*e.g. office, mails, instantaneous messages, others*), are transposed into electrical intensities, tensions, and powers.

<sup>1</sup> It leads to the definition of a multidimensional model enabling the development of new decision-making tools for work optimisation and organisation learning.

**Performance and stress at work.** A vertex  $s$  directly connected with two other vertices  $r$  and  $t$  is likened to a dipole owning a resistance  $R$ . Thereby, our work introduces the original notion of *tension* of a social network. To compute a *load-capacity* ratio of the enterprise social network, by analogy with  $Ps$  and  $Pmax$ , we use the OHM's laws:

$$U_{rt} = R_s \cdot I_{rt} \text{ and } P_s = R_s \cdot I_{rt} \cdot I_{rt}^2 = U_{rt} \cdot I_{rt}^2 / R_s = U_{rt} \cdot I_{rt},$$

where  $U_{rt}$  represents the electrical tension and  $P_s$  represents the delivered power by a vertex of which maximal admissible power is noted  $Pmax$ , with:  $U_{max} = \sqrt{R \cdot Pmax}$  and  $I_{max} = \sqrt{Pmax / R}$ .

Our goal is to introduce a *stress* at work measure. This measure uses the *Joule effect* to estimate the enterprise social network components *warm-up*<sup>2</sup> and to prevent risks of performance degradation or psychosocial trouble. According to JOULE, we defined the *warm-up* as follows:  $T \cdot \rho = W = R \cdot I^2 \cdot \Delta t$ , with  $\rho = 1$  and  $\Delta t$  time interval.

**Predictive system for human capital management.** According to the experiment of E. BRANLY about radio-conduction (1890), which has demonstrated that some flows can appear between points without visible connections, we state that significant flows of information can exist between two vertices without an edge connecting them. We apply our allegation to provide a new dynamic model of EISNA where flows are recursively depending on structures and *vice-versa*.

### 2.3 Semantic SNA

Superposition of ontologies to static or dynamic SNA adds a semantic dimension necessary to a conceptual, qualitative and faceted classification of flows and structures. Currently, to our knowledge, a few significant work has been published in this area.

J. JUNG AND J. EUZENAT make coincide social graphs, annotations and ontologies in a three-dimensional semantic SNA model, in order to build *consensual* ontologies<sup>3</sup> [7]. *Social tagging* or *reciprocal evaluation* between members of a social network shows how human interaction produces a valuation on which a reliable *degree of confidence* can be computed. Our work, dedicated to EISNA, makes a difference by overlaying ontologies on static and electrodynamic models, to exhibit knowledge-based communities and information flows.

## 3 Multidimensional Synergies in EISNA

### 3.1 Static EISNA and Cognition

We introduce a conceptual dimension to enable classification of outcomes by overlaying endogenous trades-oriented ontologies upon social graphs. Figure 1 illustrates the way we qualify static EISNA. Explicit relationships between

<sup>2</sup> An excessive warm-up produces a burn-out.

<sup>3</sup> An ontology is a formal explicit specification of a shared conceptualisation [6].

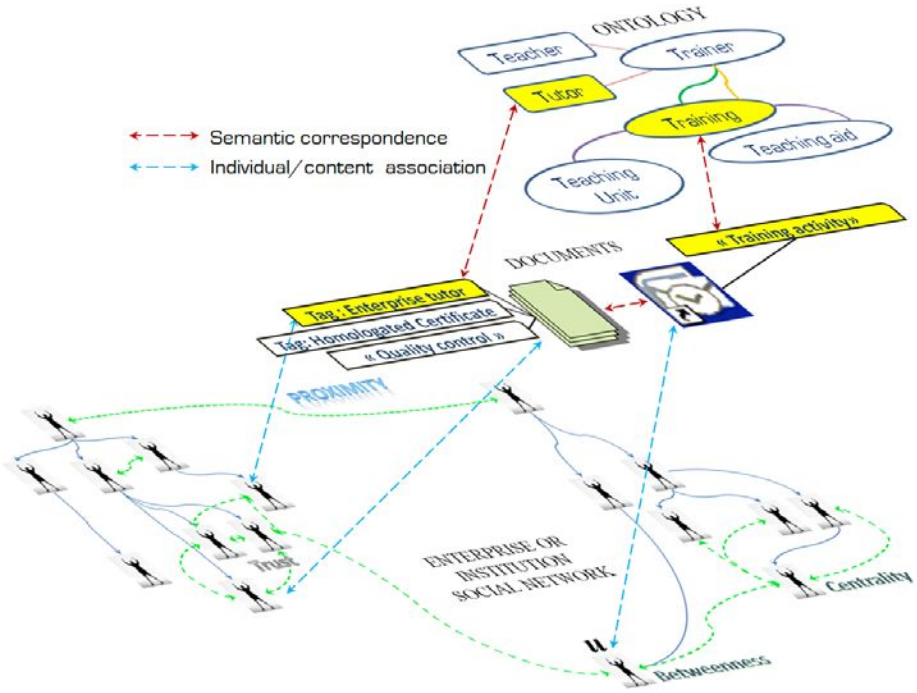


Fig. 1. Semantic betweenness on  $R_h$ ,  $R_{si}$  - eq. [3]

human resources  $R_h$ , enterprise content  $R_{si}$  and content annotations  $E_{si}$  are used to enrich EISNA and to discover some implicit relationships  $R'$ . We define an explicit relationship  $R(D, D')$  where:  $D = R_h \vee R_{si} \wedge D' = R_h \vee R_{si} \vee E_{si}$ .

We define a set of new measures by introducing a weighting ratio  $C_p$ , based on the cardinality of  $R$ .  $R$  is parametrised by  $(pD, pD')$  that are pointing out  $(D, D')$  and, optionally, by  $eD, eD'$  that are restricting  $(D, D')$ .  $C_p$  enables to declare authorized forms of  $R$  between  $R_h, R_{si}, E_{si}$  as weighting ratios in FREEMAN's social graphs measures or any other type of compatible metrics - e.g NEWMAN's betweenness<sup>4</sup>. Equation (2) formally defines  $C_p$ . Applied with  $eD \in pD \wedge eD' \in pD'$ ,  $C_p$  is used to extend the equation (1) to equation (3):

$$C_p = \frac{Card_{R(pD, pD', eD, eD')}}{SNA\ metric} \tag{2}$$

$$I_{C_p}(u) = \sum_{ij} \frac{\sigma(i, u, j)}{\sigma(i, j)} \times \frac{Card_{R(pD, pD', eD, eD')}}{\sum_{ij} \frac{\sigma(i, u, j)}{\sigma(i, j)}} \tag{3}$$

Equation (3) introduces a new measure of *semantic betweenness* based on [4]. Figure 1 illustrates this principle, with  $C_{R(pD, pD', eD, eD')} = 3$ , if  $pD = R_h, pD' =$

<sup>4</sup>  $R$  can make up a relationship hierarchy based on the pattern  $R'(pD, pD') \rightarrow R(pD, pD')$  with  $pD$  or  $pD'$  common to  $R$  and  $R'$ .

$Rsi, eD = \{u\}$  and, optionally,  $eD'$  is a singleton {"Training"}. Discovered knowledge from these conceptual associations is the strong point of this new "smart" measure.

### 3.2 Dynamic EISNA and Electrodynamic

Our hypothesis stands on electrodynamic to preview implicit social networks induced by information flows between closest neighbours. According to [5], we add our own simplification of an electromagnetic induction  $B$  with  $I$  intensity (Ampere) and  $d$ , distance (m) from the conductor to a point<sup>5</sup>. Mutual induction is an electromagnetic interaction between close electrical flows. Mutual inductance of circuits 1 and 2 is noted  $M_{1/2}$ . It is defined as follows:

$$B = \frac{\mu_0}{2\pi} \cdot \frac{I}{d} \text{ and } L = \frac{\Phi}{I} \text{ and } M_{1/2} = \frac{\Phi_2}{I_1} \tag{4}$$

We use electrodynamic laws to identify communities and to predict edges creation and deletion within the social graph. We overlay domain ontologies upon our dynamic SNA model to select socio-semantic sub-networks linked to skills, training or similar socio-professional knowledge. Our innovative model is powerful to develop new decision-making functions for human and social capital management - e.g. recommendations for training plan or teams organisation.

### 3.3 Semantic EISNA

In electrodynamic, the reactance (in Ohms) describes the energy opposed to an alternative current. In psychology, the reactance characterises a *state of negative motivation following a menace (supposed to be real) of individual freedom restriction that is translated into a influence resistance* [2]. In our work, we propose to use the *reactance*  $\Psi$  as a measure of individual stress.

We propose a first set of knowledge dedicated to the identification of individual stress, represented by the following rules and inferring the following axioms:

- \* **rule 1:**  
 If  $CC_u = \frac{charge_u}{capacity_u}$  increases and  $CC_u > 80\%$ , then  $\Psi_u$  significantly increases.
- \* **rule 2:**  
 if  $P_u = \frac{resistance_u \cdot intensity_{(e1,u,e2)^2}}{Pmax_u}$  increases and  $P_u \leq 1$ , then  $\Psi_u$  and  $warm - up_u$  increases ( $P_u$  represents a used power).
- \* **rule 2 bis** (inference learning on rule 2):  
 if  $warm - up_u$  increases, then  $\Psi_u$  increases.
- \* **rule 3:**  
 if  $P_u$  increases and  $P_u > 1$ , then  $\Psi_u$  decreases,  $Pmax_u$  decreases and  $warm - up_u$  quickly increases ( $P_u$  has exceeded  $Pmax_u$ ).
- \* **rule 3 bis** (inference learning on rule 3 and experts supervision):  
 if  $\Psi_u$  decreases and  $warm - up_u$  increases, then quick decreasing of  $Pmax_u$  and break-down risk.

---

<sup>5</sup> A point in an euclidean plan, orthogonal to the conductor.

- \* **axiom 1** (*inference supervised learning on rule 1*):  
if  $CC_u \leq 0.8$ , then risk to lose socio-professional performances.
- \* **axiom 2** (*inference learning on rule 3 and 3 bis*):  
if  $P_u > 1$ , then risk of socio-professional troubles.
- \* **axiom 3** (*inference supervised learning on axioms 1 + 2*):  
performance optimisation is equivalent to  $CC_u > 0.8$  and  $P_u \leq 1$ .
- \* **axiom 4** (*learning from symmetry on axiom 3*):  
risk of socio-professional troubles is equivalent to risk of loss of socio-professional performances.

From the equations system underlying these rules and axioms, we are currently formalising a scalar measure of reactance  $\Psi_u$ , paired with the notion of tension.

## 4 Conclusion

The purpose of our model is to integrate static, electrodynamic and semantic dimensions in EISNA. Our current proposal consists of two measures (*semantic betweenness, reactance*) and an innovative system, defined by a multidisciplinary approach. This work is a baseline for the development of new decision-making functions and tools, dedicated to socio-professional troubles risk prevention - *i.e.* detecting an individual possible burn-out before it occurs -, performances loss risk prevention - *i.e.* identifying lacks of knowledge or misused skills in teams - and social risk prevention in enterprises and institutions - *i.e.* mismatching between skills and roles leading to physical or psychological damages.

Our proposal is currently evaluated in the context of experiments related to the SOCIOPRISE project. This work is currently in progress towards the accurate integration of physics models in order to formalise *a complex and multi-dimensional model (static, dynamic and semantic) dedicated to enterprises and institutions social network analysis*.

## References

1. Barabasi, A.-L., Albert, R.: Emergence of scaling in random networks. *Science Magazine* 286(5439), 509–512 (1999)
2. Brehm, J.W.: *A Theory of Psychological Reactance*. Academic Press, London (1966)
3. Erdős, P., Rényi, A.: On random graphs. *Publicationes Mathematicae* 6, 290–297 (1959)
4. Freeman, L.C.: A set of measures of centrality based on betweenness. *Sociometry* 40, 35–41 (1977)
5. Gardiol, F.: *Electromagnétisme*. Dunod (1987)
6. Gruber, T.R.: Toward principles for the design of ontologies used for knowledge sharing. *International Journal of Human Computer Studies* 43(5/6), 907–928 (1995)
7. Jung, J., Euzenat, J.: Towards semantic social networks. In: Franconi, E., Kifer, M., May, W. (eds.) *ESWC 2007*. LNCS, vol. 4519, pp. 267–280. Springer, Heidelberg (2007)
8. Sowa, J.F.: *Knowledge Representation: Logical, Philosophical, and Computational Foundations*. Brooks Cole Publishing Co., Pacific Grove (2000)

# Knowledge Based Systems and Metacognition in Radar

Gerard T. Capraro<sup>1</sup> and Michael C. Wicks<sup>2</sup>

<sup>1</sup> Capraro Technologies, Inc.,  
2118 Beechgrove Place, Utica, NY 13501 USA  
GCapraro@CapraroTechnologies.com  
<sup>2</sup> Air Force Research Laboratory Sensors Directorate,  
26 Electronic Parkway Rome, NY 13441 USA  
Michael.Wicks@rl.af.mil

**Abstract.** An airborne ground looking radar sensor's performance may be enhanced by selecting algorithms adaptively as the environment changes. A short description of an airborne intelligent radar system (AIRS) is presented with a description of the knowledge based filter and detection portions. A second level of artificial intelligence (AI) processing is presented that monitors, tests, and learns how to improve and control the first level. This approach is based upon metacognition, a way forward for developing knowledge based systems.

**Keywords:** Metacognition, radar, signal processing, CFAR, knowledge-based, sensors.

## 1 Introduction

The desire to anticipate, find, fix, track, target, engage, and assess, anything, anytime, anywhere (AF2T2EA4) by the US Air Force (USAF) will require changes to how we modify, build, and deploy radar and sensor systems. The US Air Force Research Laboratory (AFRL) is attacking these issues from a sensor and information perspective and has generated a new sensing concept in their defining of layered sensing [1].

How can the US Air Force system of the future detect and identify threats and meet the implicit requirements of this scenario in a timely manner? We must, as a first step to full automation, implement the following ground breaking changes: place more compute intensive resources closer to sources of the information gathering – e.g. assign tasks to sensors to look for “triggers” created from intelligence surveillance and reconnaissance (ISR) sources, provide for the analysis of intelligence data automatically and without human involvement, move the human sensor operator from managing data - to managing actionable knowledge and sensor aggregation, and develop these “triggers” and rules for automatic assignment and management of heterogeneous sensors to meet dynamic and abstract requirements.

Sensor performance may be enhanced by selecting algorithms adaptively as the environment changes. It has been shown [2-12], that if an airborne radar system uses prior knowledge concerning certain features of the earth (e.g. land-sea interfaces) intelligently, then performance in the filtering, detection and tracking stages of a radar processing chain improves dramatically.

One design of an intelligent radar system that processes information from the filter, detector, and tracker stages of a surveillance radar, investigated by the USAF and under the KASSPER program, was specifically designed for an Airborne Intelligent Radar System (AIRS). Futuristic advanced intelligent radar systems will cooperatively perform signal and data processing within and between sensors and communications systems while utilizing waveform diversity and performing multi-sensor processing, for reconnaissance, surveillance, imaging and communications within the same radar system. A high level description of AIRS is shown in Figure 1 and is described in detail, [8, 11], in the literature.

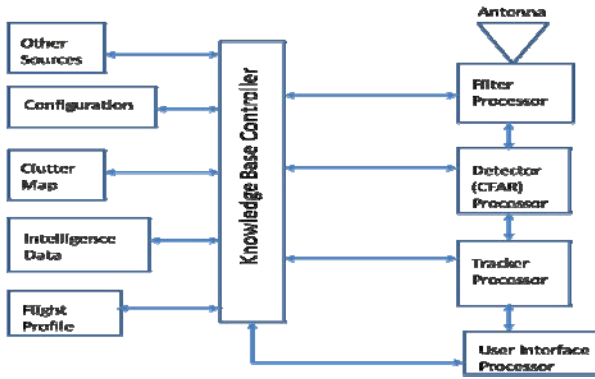


Fig. 1. Airborne intelligent radar system

## 2 Description of AIRS

The knowledge based (KB) signal and data processing architecture shown in Figure 1 represents one radar sensor system. The major components in the figure are labeled as processors with the knowledge base controller as the major integrator for communications and control of the individual processors. Data about the radar, its frequency of operation, antenna configuration, where it is located on the aircraft, etc. is provided by the block labeled Configuration in Figure 1. Map data is loaded before each mission for estimating clutter returns and for registering its location relative to the Earth and with other sensor platforms. It is also preloaded with flight profile data and is updated continuously from the platform's navigation system. It also will receive information from the intelligence community both before a mission and throughout the mission. During flight, the KB Controller (KBC) will receive information about weather, jammer locations, requests for information, discrete locations, fusion information, etc. The radar system is assumed to be aboard a surveillance aircraft flying a known and repeatable path over the same terrain, e.g. a "racetrack" flight path. Therefore it can learn by monitoring the performance of different algorithms over repeatable passes over the same terrain.

The KBC performs the overall control functions of AIRS. It assigns tasks to all processors, communicates with outside system resources, and "optimizes" the system's global performance. Each individual processor "optimizes" its individual performance



measures, e.g. signal-to-noise ratio and probability of detection. The tracker with the KBC, for example, "optimizes" the number of correct target tracks and "minimizes" the number of missed targets, incorrect tracks, and lost tracks. The KBC handles all interrupts from the User Interface Processor, assigns tasks to individual processors based upon user requested jobs, generates information gathered from sources to enhance the performance of the individual processors, works with other sensors and outside sources for target identification, and provides the User Interface Processor periodic and aperiodic data for answering queries and requests from the user.

### 3 Metacognition

Metacognition is a term used in educational psychology. It refers to the active control over the cognitive processes one uses to understand how they learn. The process of planning for a given learning task and monitoring and evaluation of one's comprehension are all metacognitive tasks. According to [13] a true AI system must be able to work at two levels: the knowledge and the metaknowledge level. Here the metaknowledge level is the "how to think" level, i.e. "To give knowledge is like giving a fish to a starving man, to give metaknowledge is like teaching him how to fish".

Metacognition according to [14] contains multiple variables. There is the person variable dealing with an individual and their capabilities e.g. a radar expert knows more about signal processing algorithms than the average electrical engineer. Then there are task variables which refers to the type of mental activity e.g. it is easier in general for a radar engineer to solve a simple mathematics problem than an organic chemistry problem. Third there are strategy variables that relate to the alternative approaches to a mental task, e.g. to understand the contents within a chapter of a mathematics book, it is a good idea to solve all the problems at the end of the chapter or reread the chapter the night before a test. Some experts include a self-monitoring component to metacognition or a way to measure one's understanding, e.g. where one evaluates their levels of comprehension and mental performance.

The AI community has been studying metacognition for many years and attempting to understand how we can provide metaknowledge, metamemory, metareasoning, databases, and knowledge bases and thereby have our AI systems work at multiple levels so that it can learn new facts, rules, and strategies especially under dynamic environments. Researchers have been working on how to construct algorithms and software tools [15] that will be able to learn. Some researchers believe that analogy is the core or the center of how we think [16].

### 4 KB Filtering and CFAR

Doppler spectrum occupied by clutter, reduces the effectiveness of Airborne Moving Target Indicator (AMTI). Also, Fast Fourier Transform (FFT)-based Doppler filtering is suboptimum because the clutter returns are no longer confined to the zero Hertz filter. Platform motion and sidelobe returns broaden the clutter spectrum, spreading clutter energy into adjacent Doppler bins. This further complicates detection processing. It is in situations such as this that the use of a single combination of filtering and Constant False Alarm Rate (CFAR) algorithms will produce excessive false alarms,

because it cannot be designed to be optimum for each and every scenario to which it must be applied. In light of the many constraints imposed upon radar systems, improvements in detection performance are most likely to be a result of advanced processing techniques able to recognize the existence of these situations and apply appropriate processing while effectively maintaining a constant false alarm rate and an adequate detection probability. Consider Figure 2 and Figure 3.

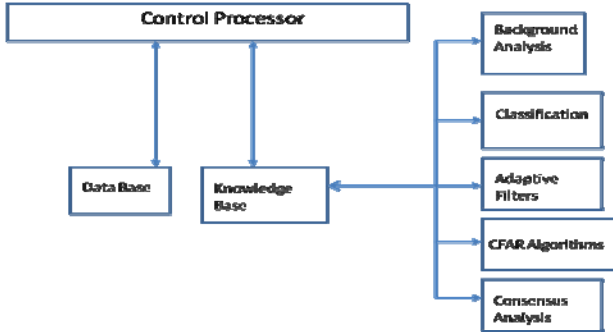


Fig. 2. Filtering, CFAR, and Control Processor

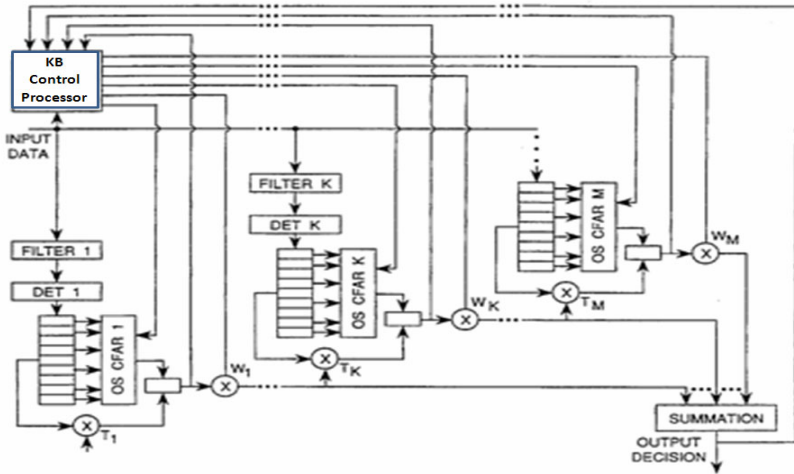


Fig. 3. Control Processor

## 5 Performance Processor

The control processor performs filtering and detection by setting the weights and summing the results for an assigned filter/CFAR pair. The choice of those pairs and the management of the control processor is accomplished by the performance processor or the metacognition portion of the KB Controller. Figure 4 provides an overview of the performance processor and its interaction with the control processor.

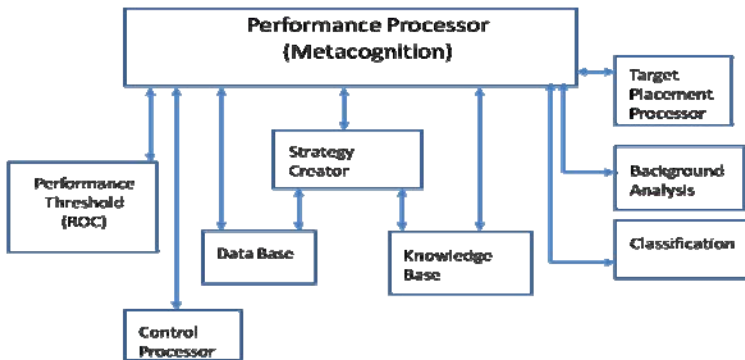


Fig. 4. Performance Processor

To monitor performance the performance processor will place synthetic targets with “known” radar response levels within the different terrain type areas and multiple synthetic targets at varying boundaries of the different terrain types. This will allow the performance processor an opportunity to assess the different strategies for different terrain types. The synthetic targets are not known by the control processor and once detected the performance processor will excise them before they are accessed by the track processor. By applying these synthetic targets the performance processor can obtain a realistic estimate of the probability of false alarm and the probability of detection as it relates to the environment and the chosen filter/CFAR algorithms, the weights chosen, and the strategy.

## 6 Conclusions

It is our belief that the techniques expressed here are just the beginning of how metacognition along with knowledge based approaches will allow the radar and sensor community, in general, to expedite the development of truly intelligent sensor systems.

## References

1. Bryant, M., Johnson, P., Kent, B.M., Nowak, M., Rogers, S.: Layered Sensing Its Definition, Attributes, and Guiding Principles for AFRL Strategic Technology Development, <http://www.wpafb.af.mil/shared/media/document/AFD-080820-005.pdf>
2. Baldygo, W., Wicks, M., Brown, R., Antonik, P., Capraro, G., Hennington, L.: Artificial intelligence applications to constant false alarm rate (CFAR) processing. In: Proceedings of the IEEE 1993 National Radar Conference, Boston, MA, pp. 275–280 (1993)
3. Senn, R.: Knowledge Base Applications To Adaptive Space-Time Processing, AFRL-SN-TR-146, Final Technical Report (July 2001)
4. Antonik, P., Shuman, H., Li, P., Melvin, W., Wicks, M.: Knowledge-Based Space-Time Adaptive Processing. In: Proceedings of the IEEE 1997 National Radar Conference, Syracuse, NY, pp. 372–377 (1997)

5. Wicks, M.C., Baldygo, W.J., Brown, R.D.: Expert System Constant False Alarm Rate (CFAR) Processor, U. S. Pat. 5,499,030 (1996)
6. Multi-Channel Airborne Radar Measurement (MCARM) Final Report, Volume 1 of 4, MCARM Flight Test, Contract F30602-92-C-0161, for Rome Laboratory/USAF, by Westinghouse Electronic Systems
7. Capraro, C.T., Capraro, G.T., Weiner, D.D., Wicks, M.: Knowledge Based Map Space Time Adaptive Processing (KBMapSTAP). In: Proceedings of the 2001 International Conference on Imaging Science, Systems, and Technology, Las Vegas, Nevada, pp. 533–538 (2001)
8. Farina, A., Griffiths, H., Capraro, G., Wicks, M.: Knowledge-Based Radar Signal & Data Processing. NATO RTO Lecture Series, vol. 233 (2003)
9. Capraro, C.T., Capraro, G.T., Bradaric, I., Weiner, D.D., Wicks, M.C., Baldygo, W.J.: Implementing Digital Terrain Data in Knowledge-Aided Space-Time Adaptive Processing. *IEEE Trans. on Aerospace and Electronic Systems* 42(3), 1080–1099 (2006)
10. Capraro, C.T., Capraro, G.T., Wicks, M.C.: Knowledge Aided Detection and Tracking. In: Proceedings of the IEEE 2007 National Radar Conference, Boston, MA, pp. 352–356 (2007)
11. Capraro, G., Wicks, M.: An Airborne Intelligent Radar System. In: Radar 2004, International Conference on Radar Systems, Toulouse, France (2004)
12. Melvin, W.L., Wicks, M.C., Chen, P.: Nonhomogeneity Detection Method and Apparatus for Improved Adaptive Signal Processing, U. S. Pat. 5,706,013 (1998)
13. Pitrat, J.: AI Systems Are Dumb Because AI Researchers Are Too Clever. *ACM Computing Surveys* 27(3), 349–350 (1995)
14. Cox, T.M.: Metacognition in Computation: A Selected History. In: AAI Spring Symposium (2005), <http://www.aaai.org/Papers/Symposia/Spring/2005/SS-05-04/SS05-04-002.pdf>
15. Shapiro, S.C., Rapaport, W.J., Kandefer, M., Johnson, F.L., Goldfain, A.: Metacognition in SNePS. *AI Magazine* 28(1), 17–31 (2007)
16. Hofstadter, D.: Analogy as the Core of Cognition, <http://prelectur.stanford.edu/lecturers/hofstadter/analogy.html>

# Maximus-AI: Using Elman Neural Networks for Implementing a SLMR Trading Strategy

Nuno C. Marques and Carlos Gomes

CENTRIA — Departamento de Informática, Faculdade de Ciências e Tecnologia,  
Universidade Nova de Lisboa and GoBusiness  
maximus-ai@gobusiness.pt

**Abstract.** This paper presents a *stop-loss - maximum return* (SLMR) trading strategy based on improving the classic moving average technical indicator with neural networks. We propose an improvement in the efficiency of the long term moving average by using the limited recursion in Elman Neural Networks, jointly with hybrid neuro-symbolic neural network, while still fully keeping all the learning capabilities of non-recursive parts of the network. Simulations using Eurostoxx50 financial index will illustrate the potential of such a strategy for avoiding negative asset returns and decreasing the investment risk.

## 1 Introduction

Several authors (e.g. [3], [11], [1]) show an empirical evidence confirming that Normal Distribution doesn't fit the behaviour of the financial assets returns and that leads to risk underestimation. This risk underestimation increases the probability of financial crises. We will extend a *stop-loss - maximum return* (SLMR) trading strategy [1], assuming implicit time dependency in financial assets and using technical indicators, namely moving averages, as a risk reducing device. Traditional moving averages, taken over a fixed number of days, are too strict, and cannot adapt to different market conditions. This paper presents a model that can dynamically adjust the number of days to be considered for calculating a moving average according with the patterns observed over a given set of fundamental and technical market conditions. The goal is to have a descriptive approach, that can be used jointly with the SLMR model. So, this approach is different from the more direct trend of most previous works applying neural networks to stock market price prediction.

Next section conjoins the models presented in [10] and [1], for encoding knowledge in Elman Neural Networks in the case of financial markets. Then section 3 will validate such a strategy by using the DJ *Eurostoxx50* financial index for illustrating the potential of such a strategy. Finally some conclusions will be drawn.

## 2 Representing the Intelligent Moving Average for Training Elman Neural Networks

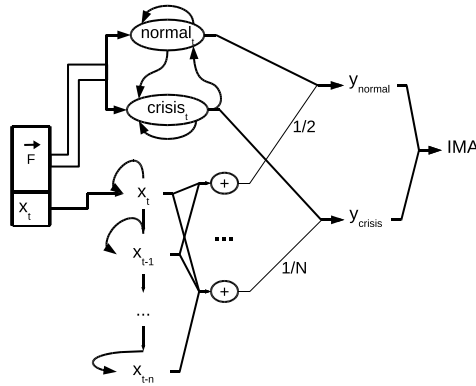
Elman neural networks make a direct extension of the back-propagation learning algorithm, by extending the network with recurrent context units (or memory units) [4]. The

recurrent weights have a fixed value (usually 1), so the back-propagation algorithm still can be applied to this class of networks. This variant of the backpropagation algorithm is known as backpropagation through time [4]. The main problem with Elman neural networks (as with other kinds of neural networks) is that they have too generic learning capabilities: instead of representing logical functions (as in initial proposal by [9]), the generic learning capabilities of neural networks made them black boxes. Also, the back-propagation algorithm distributes learning for a given pattern by as many units as possible, while errors in model output are minimised regarding the target value provided in the dataset. Unfortunately there are only generic a priori specifications regarding what are the best proper input encodings and neural architectures for a given problem. E.g. Fuzzy Neural Networks (e.g. [5]) can encode *linguistic rules* inside a neural network, however the semantic of such a model is limited and the neural network computation is usually replaced by a fuzzy classifier (e.g. the recursive approach of the Elman network is harder to implement). Unguided learning in neural networks is also sensible to early convergence in local minima and unsuitable generalisation of the train data (overfitting). We use an alternative based on the use of hybrid neuro-symbolic methods ([7], [6]). Hybrid neuro-symbolic methods can encode logical programs in a feed-forward neural network core (the model is usually called *the core method*). In this method the neural network keeps all its computational power while still encoding both statistical and logic models. The system also learns to adjust itself (i.e. fits) to experimental data. Moreover, in [2] the authors present experimental evidence that backpropagation learning can be guided by encoding logical rules. [8] shows that logical knowledge may not be enough for different classes of problems and generalises the represented knowledge from *true – false* values to real values. Finally, by explicitly modelling the time series inside the network, [10], shows how such models can be used for computing intelligent moving averages. Here these results will be applied to our new SLMR/Maximus proposal of a new, and safer, trading strategy.

**SLMR model.** To evaluate the risk underestimation, five indicators were created: 1) Serial Correlation; 2) Fama Multiples; 3) Correlation Breakdown; 4) Moving Average; 5) Trading Range Break-Out. Combining the information from the five indicators of risk underestimation the Stop Loss -Maximum Return (SLMR) investment strategy was suggested to achieve superior returns, statistically different, without risk increase and exposure to rare events [1]. The SLMR Strategy resumes as follows:

- **The Stop Loss** — its function is to suggest signals to sell the financial asset to avoid the huge losses following risk underestimation;
- **The Maximum Return** — its role is to suggest signals to buy the financial asset, when it shows strong signals of recovering, after the huge fall had happened.

**Using an Intelligent Moving Average.** Diagram 1, shows the specification for implementing the calculation of the intelligent moving average function taking into account the SLMR model. The moving average component is encoded in the lower part of the graph and follows the encoding studied in [10]: each node can be associated with a state or variable. Each arrow represents the memorisation of a given (previous state) value by other state. When a transition has a number associated to it, this memorisation should be



**Fig. 1.** Diagram of proposed model for implementing the financial intelligent moving average

multiplied by previous state value. In the diagram the addition corresponds to the sum of all connected variables (and is not a variable). The upper part of the diagram is reserved for the SLMR model. The two main states *crisis/Stop Loss* and *normal/Maximum Return* are represented. These states are activated by current financial indicators ( $F$ ). Although we don't know how many days should be used for computing the moving average under each market condition (as defined by  $F$ ), we know we should depart from extreme values, i.e. a 40 day *short* moving average during *normal* periods ( $y_{normal}$ ) and a one year moving average ( $y_{crisis}$ ), during *crisis* periods.

We use the semi-recursive Elman network [4] where the initial state in the diagram will represent the input value for the neural network (i.e. the value of the time series vector  $x_t$ ). Input values are connected to the first layer (in the graph), represented by the hidden unit layer and the contextual (Elman) layer. Each transition among a unit  $x_t \rightarrow x_{t-1}$  will represent a recursive connection from  $x_t$  to a memory contextual unit  $x_{t-1}$ , followed by a feed forward connection (without further information we will assume weights of 1.0) to the hidden layer unit  $x_{t-1}$ . The connection from the hidden layer to the output layer always uses the value  $1/N$  (i.e. a multiplying factor). The states *crisis* and *normal* were implemented by two hidden neurons and correspondent contextual units in an Elman neural network. Since we still use the basic specification for the core method [7], *true* and *false* values can be encoded by 1.0 or  $-1.0$ , respectively. Due to the use of the hyperbolic tangent as an activation function, these two boolean indicators can be connected to a second layer of hidden neurons ( $y_{norm}$  and  $y_{crisis}$ ) [8] while the memory effect is achieved by using the contextual units in the Elman neural network. For conjoin the non-linearity with the linear moving average calculation all values for  $x_t$  are normalized to the range  $[-0.1, 0.1]$  [10]. As a result, when activated, the  $-1/1$  output will switch off a  $[-0.5, 0.5]$  value of the moving average (or, conjoined with neuron bias, will accept it). This way, the network will compute the most appropriate moving average as a result of conjoining both  $y_{state}$  values. All remaining feed-forward weights are randomly set in a the  $[-0.1, 0.1]$  range (an usual procedure if we don't want to encode a priori knowledge in the model). Finally a small amount of noise ( $+ - 0.01$  or 2% of neural weights) was applied to all connection weights.

### 3 Predicting the Best Moving Average in SX5E

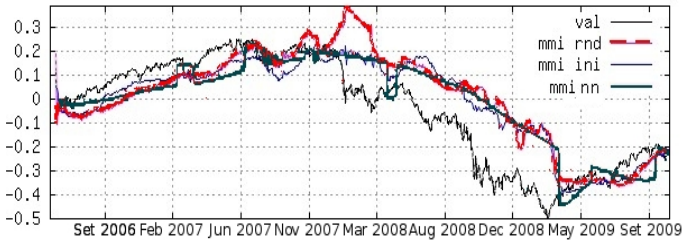
The network just described was tested over the European *DJ EURO STOXX 50* financial index (*SX5E*). Daily prices in the 9 year period of 1 September 2000 to 9 November 2009 were considered as our data. *F* vector (in diagram 1) was set to represent several fundamental measures provided by our financial partner (e.g. the *German CDS index*, the *US CPI*, *urban consumers SA index*, *US Generic Government 1 Month Yield* or even *gold spot*, *s/OZ commodity*) and technical measures (e.g. daily and other average variations and other more evolved risk measures). A precalculated moving average was precalculated over *SX5E* and was set as the *x* value in diagram 1. Validation data was always set to the period starting in 14 June 2006 (selected for including the 2008 financial crisis, but also to give some positive return period). Please notice that this period has a  $\frac{2822.72-3414.21}{2822.72} = -17.32\%$  negative return.

The network was trained using the following strategy: a *premonition* factor was set to 20 trading days. When *SX5E* index has a positive variation in the next 20 trading days *Ima* training target is set as to a short (40 days) moving average (this will advise our financial simulator to give buying instructions). On the other hand, if *SX5E* has a negative variation, a more conservative one year *Ima* is set as the target value. This premonitive train *Ima* value is plotted against *SX5E* index in figure 2 – A. This represents all available data. All values were normalised to values in the  $[-0.1, 0.1]$  interval (for enabling regression while still using the hyperbolic tangent activation function).

Figure A - SX5E Dataset



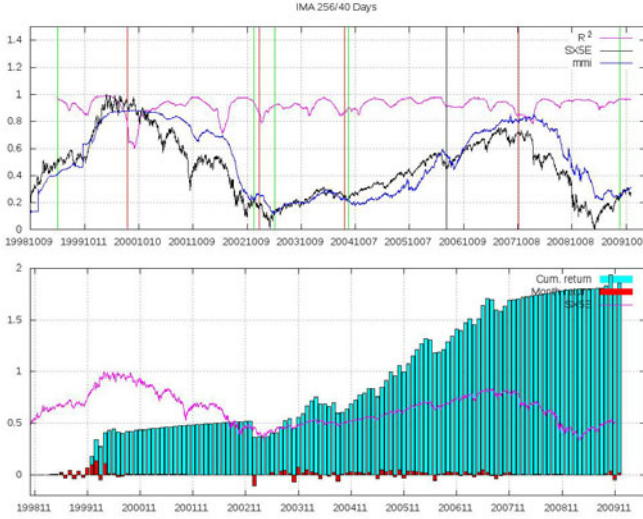
Figure B - SX5E Testset



**Fig. 2.** Dataset and results over validation data for the intelligent moving average (description in text)



As in [10], we can notice (figure 2 – B) that the initialised Elman network is much more predictable and stable than the randomly initialised one. Indeed output values are always near to the target *Ima* and the sharp effects of input index variations are much less noticed in this network output. Also quite good results can be noticed regarding comparison with the targeted premonitive *Ima*. This seems to point that during validation the neural network (without receiving any premonition information) knows how to use the *F* vector to predict the future<sup>1</sup>.



**Fig. 3.** Dataset and results over training and validation data for the intelligent moving average (description in text)

For a more quantitative measure, a financial simulator was developed. Neural network *Ima* was used jointly with an  $R^2$  goodness of fit test indicator for measuring instability. The non-invested periods have return given by the Euribor rate. Figure 3 plots normalised values of the technical indexes and simulator financial decisions (represented as vertical lines, green for buying decisions and red for selling decisions) over *IMA* (initialized for 40 to 256 trading days). A black vertical line identifies the start of the test data. The bar chart presents monthly and cumulative returns after each decision. Under this conditions global return during test period of this *Ima* index was 58%, while moving averages, *MA256* return was 43% and *MA40* had a negative return of -37%. Comparisons were also made with several major fund houses, using the same benchmark as SLMR (DJ EuroStoxx50) in this period: all of them lost value in this period (from -12% to -28%).

<sup>1</sup> Notice this will always be a risky prediction: selected features (i.e. *F*) may be relevant only to this financial crisis so, as usual, *past incomes can not be used to predict future ones*.

## 4 Conclusions

Humans are not that good trying to optimise decisions, e.g.: should we use a 40 day moving average or a 56 day moving average for short selling periods, or what would be the best strategy to combine our fundamental and technical parameters. With the proposed method we have illustrated how we can achieve such optimizations by joining a SLMR trading strategy with a moving average calculation inside an Elman Neural Network [4]. This is very useful if we want to leave some information under-specified for our method, but would also be very useful in conjoining the different technical indicators that are now being used for SLMR. Moreover, backpropagation through time algorithm [4] was used to discover unspecified trading patterns in the dataset. However that search was guided, in the sense that we always want some kind of index or moving average. In fact, we need an indicator that can be studied regarding its performance and rating by our financial experts, not an opaque classification decision. So, the acquired network can be seen as a non-parametric statistical model of the proposed MMI.

The two states used for the SLMR model have a logical interpretation, and the inclusion of further knowledge can be done by using the core method or its extensions (e.g. [7] and [8]). For this study, the only a priori assumption was the use of different moving averages for a given (changeable) period of time: not only encoding more logical knowledge would be outside the scope of this paper, but also a no-knowledge approach is more appropriate to validate the core *Ima-NN* extension to the financial data. Future work will address the conjunction and automatic optimization of economic models for reasoning (based on fundamental and technical features) and on what is the most probable short term economic scenario.

## References

1. Gomes, C.: Maximus investment fund, Tech. report, GoBusiness (2010)
2. Bader, S., Hölldobler, S., Marques, N.: Guiding backprop by inserting rules. In: ECAI 2008 Workshop on Neural-Symbolic Learning and Reasoning, Greece, vol. 366, CEUR (2008)
3. Brock, W., Lebaron, B., Lakonishok, J.: Simple technical rules and stochastic properties of stock returns. *Journal of Finance* 47, 1731–1764 (1992)
4. Elman, J.L.: Finding structure in time. *Cognitive Science* 14, 179–211 (1990)
5. Feuring, T.: Learning in fuzzy neural networks. In: Proc. IEEE Int. Conf. Neural Networks, pp. 1061–1066 (1996)
6. d'Avila Garcez, A.S., Broda, K.B., Gabbay, D.M.: Neural-Symbolic Learning Systems — Foundations and Applications. In: Perspectives in Neural Computing, Springer, Berlin (2002)
7. Hölldobler, S., Kalinke, Y.: Towards a massively parallel computational model for logic programming. In: ECAI 1994 Workshop on Combining Symbolic and Connectionist Processing, pp. 68–77 (1994)
8. Marques, N.C.: An extension of the core method for continuous values: Learning with probabilities. In: New Trends in Artificial Intelligence, pp. 319–328. APPIA (2009)
9. McCulloch, W.S., Pitts, W.: A logical calculus of the ideas immanent in nervous activity. *Bulletin of Mathematical Biophysics* 5, 115–133 (1943)
10. Marques, N., Gomes, C.: T.: An intelligent moving average. In: Proceedings of the 19th European Conference on Artificial Intelligence - ECAI 2010 (2010)
11. Sheikh, A.Z., Quiao, H.: Non-normality of market returns. *The Journal of Alternative Investments* 12(3) (2009)

# A Formalism for Causal Explanations with an Answer Set Programming Translation

Yves Moinard

INRIA Bretagne Atlantique, IRISA, Campus de Beaulieu,  
35042 Rennes cedex, France  
moinard@irisa.fr

**Abstract.** We examine the practicality for a user of using Answer Set Programming (ASP) for representing logical formalisms. Our example is a formalism aiming at capturing causal explanations from causal information. We show the naturalness and relative efficiency of this translation job. We are interested in the ease for writing an ASP program. Limitations of the earlier systems made that in practice, the “declarative aspect” was more theoretical than practical. We show how recent improvements in working ASP systems facilitate the translation.

## 1 Introduction

We consider a formalism designed in collaboration with Philippe Besnard and Marie-Odile Cordier, aiming at a logical formalization of explanations from causal and “is-a” statements. Given some information such as “fire causes smoke” and “grey smoke is a smoke”, if “grey smoke” is established, we want to infer that “fire” is a (tentative) explanation for this fact. The formalization [2] is expressed in terms of *rules* such as “if  $\alpha$  causes  $\beta$  and  $\gamma$  isa  $\beta$ , then  $\alpha$  explains  $\gamma$  provided  $\{\alpha, \gamma\}$  is possible”. This concerns looking for paths in a graph and ASP is good for this. There exists efficient systems, such as DLV [6] or clasp[D] ([www.dbai.tuwien.ac.at/proj/dlv/](http://www.dbai.tuwien.ac.at/proj/dlv/) or [potassco.sourceforge.net/](http://potassco.sourceforge.net/)).

Transforming formal rules into an ASP program is easy. ASP should then be an interesting tool for researchers when designing a new theoretical formalization as the one examined here. When defining a theory, ASP programs should help examining a great number of middle sized examples. Then if middle sized programs could work, a few more optimization techniques could make real sized examples work with the final theory.

In fact, even if ASP allows such direct and efficient translation, a few problems complicate the task. The poor data types available in pure ASP systems is a real drawback, since our rules involve sets. Part of this difficulty comes from a second drawback: In ASP, it is hard to reuse portions of a program. Similar rules should be written again, in a different way. Also, “brave” or “cautious” reasoning is generally not allowed (except with respect to precise “queries”). In ASP, “the problem is the program” and the “solution” consists in one or several sets of atoms satisfying the problem. Each such set is an *answer set*. Brave and cautious solutions mean to look for atoms true respectively in some or in all the answer sets.

This complicates the use of ASP: any modification becomes complex. However, things are evolving, e.g. DLV-Complex ([www.mat.unical.it/dlv-complex](http://www.mat.unical.it/dlv-complex)), deals with the data structure problem and DLT [3] allows the use of “templates”, convenient for reusing part of a program. We present the explanation formalism, then its ASP translation in DLV-Complex, and we conclude by a few reasonable expectations about the future ASP systems which could help a final user.

## 2 The Causal Explanation Formalism

### 2.1 Preliminaries (propositional version, cf [2] for the full formalism)

We distinguish various types of statements:

*C*: A theory expressing causal statements. E.g. *On\_alarm causes Heard\_bell*.

*O*: *IS-A* links between items which can appear in a causal statement. E.g.,  
*Temperature\_39*  $\rightarrow_{IS-A}$  *Fever\_Temperature*,  
*Heard\_soft\_bell*  $\rightarrow_{IS-A}$  *Heard\_bell*.

*W*: A classical propositional theory expressing truths (incompatible facts, co-occurring facts, ...). E.g., *Heard\_soft\_bell*  $\rightarrow \neg$ *Heard\_loud\_bell*.

Propositional symbols denote states of affairs, which can be “facts” or “events” such as *Fever\_Temperature* or *On\_alarm*. The causal statements express causal relations between facts or events.

Some care is necessary when providing these causal and ontological atoms. If “*Flu causes Fever\_Temperature*”, we conclude *Flu explains Temperature\_39* from *Temperature\_39*  $\rightarrow_{IS-A}$  *Fever\_Temperature*, but we cannot state *Flu causes Temperature\_39*: the causal information must be “on the right level”.

The formal system infers formulas denoting explanations from  $C \cup O \cup W$ . The *IS-A* atoms express knowledge necessary to infer explanations. In the following,  $\alpha, \beta, \dots$  denote the propositional atoms and  $\Phi, \Psi, \dots$  denote sets thereof.

#### Atoms

1. *Propositional atoms*:  $\alpha, \beta, \dots$
2. *Causal atoms*:  $\alpha$  causes  $\beta$ .
3. *Ontological atoms*:  $\alpha \rightarrow_{IS-A} \beta$ . Reads:  $\alpha$  is a  $\beta$ .
4. *Explanation atoms*:  $\alpha$  explains  $\beta$  bec\_poss  $\Phi$ . Reads:  $\alpha$  is an explanation for  $\beta$  because  $\Phi$  is possible.

#### Formulas

1. *Propositional formulas*: Boolean combinations of propositional atoms.
2. *Causal formulas*: Boolean combinations of causal or propositional atoms.

The premises  $C \cup O \cup W$  consist of propositional and causal formulas, and ontological *atoms* (no ontological formula), without explanation atom.

#### 1. Properties of the causal operator

(a) *Entailing [standard] implication*: If  $\alpha$  causes  $\beta$ , then  $\alpha \rightarrow \beta$ .

#### 2. Properties of the ontological operator

(a) *Entailing implication*: If  $\alpha \rightarrow_{IS-A} \beta$ , then  $\alpha \rightarrow \beta$ .

(b) *Transitivity*: If  $a \rightarrow_{IS-A} b$  and  $b \rightarrow_{IS-A} c$ , then  $a \rightarrow_{IS-A} c$ .

(c) *Reflexivity*:  $c \rightarrow_{IS-A} c$ . (unconventional, keeps the number of rules low).

### 2.2 The Formal System

1. **Causal atoms entail implication:**  $(\alpha \text{ causes } \beta) \rightarrow (\alpha \rightarrow \beta)$ .
2. **Ontological atoms**
  - (a) entail implication: If  $\beta \rightarrow_{IS-A} \gamma$  then  $\beta \rightarrow \gamma$ .
  - (b) transitivity: If  $\alpha \rightarrow_{IS-A} \beta$  and  $\beta \rightarrow_{IS-A} \gamma$  then  $\alpha \rightarrow_{IS-A} \gamma$ .
  - (c) reflexivity:  $\alpha \rightarrow_{IS-A} \alpha$
3. **Generating the explanation atoms**
  - (a) *Initial case* If  $\delta \rightarrow_{IS-A} \beta$ ,  $\delta \rightarrow_{IS-A} \gamma$ , and  $W \not\models \neg(\alpha \wedge \delta)$ , then  $(\alpha \text{ causes } \beta) \rightarrow \alpha \text{ explains } \gamma \text{ bec\_poss } \{\alpha, \delta\}$ .
  - (b) *Transitivity (gathering the conditions)* If  $W \not\models \neg \bigwedge (\Phi \cup \Psi)$ , then  $(\alpha \text{ explains } \beta \text{ bec\_poss } \Phi \wedge \beta \text{ explains } \gamma \text{ bec\_poss } \Psi) \rightarrow \alpha \text{ explains } \gamma \text{ bec\_poss } (\Phi \cup \Psi)$ .
  - (c) *Simplification of the set of conditions* If  $W \models \bigwedge \Phi \rightarrow \bigvee_{i=1}^n \bigwedge \Phi_i$ , then  $\bigwedge_{i \in \{1, \dots, n\}} \alpha \text{ explains } \beta \text{ bec\_poss } (\Phi_i \cup \Phi) \rightarrow \alpha \text{ explains } \beta \text{ bec\_poss } \Phi$ .

The elementary “initial case” applies (2c) upon (3a) where  $\beta = \gamma = \delta$ , together with a simplification (3c) since  $\alpha \rightarrow \beta$  here, getting:

If  $\alpha \text{ causes } \beta$  and  $W \not\models \neg \alpha$  then  $\alpha \text{ explains } \beta \text{ bec\_poss } \{\alpha\}$ .

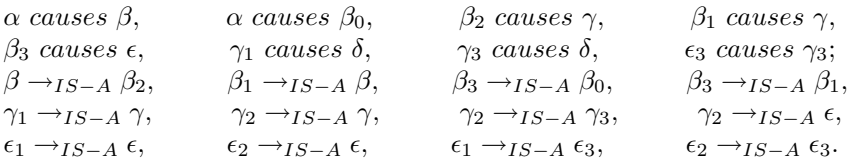
These rules are intended as a compromise between expressive power, naturalness of description and relatively efficiency.

Transitivity of *explanations* occurs (gathering conditions). The simplification rule (3c) is powerful and costly, so the ASP translation implements the following weaker rule (also, it never removes  $\{\alpha\}$  from  $\Phi$ ): [3c'] If  $W \models \bigwedge \Phi - \{\varphi\} \rightarrow \bigwedge \Phi$ , and  $\alpha \text{ explains } \beta \text{ bec\_poss } \Phi$  then  $\alpha \text{ explains } \beta \text{ bec\_poss } \Phi - \{\varphi\}$ .

An atom  $\alpha \text{ explains } \beta \text{ bec\_poss } \Phi$  is *optimal* if there is no explanation atom  $\alpha \text{ explains } \beta \text{ bec\_poss } \Psi$  where  $W \models \bigwedge \Psi \rightarrow \bigwedge \Phi$  and not conversely. Keeping only these weakest sets of conditions is useful when the derivation is made only thanks to the part of  $W$  coming from Points 1 and 2a above. This keeps all the relevant explanation atoms and is easier to read.

### 2.3 A Generic Diagram

The following diagram summarizes many patterns of inferred explanations:



This example shows various different “explaining paths” from a few given causal and ontological atoms. As a first “explaining path” from  $\alpha$  to  $\delta$  we get successively (path (1a):  $\alpha \text{ explains } \beta_2 \text{ bec\_poss } \{\alpha\}$ ,  $\alpha \text{ explains } \gamma_1 \text{ bec\_poss } \{\alpha, \gamma_1\}$ , and  $\alpha \text{ explains } \delta \text{ bec\_poss } \{\alpha, \gamma_1\}$ , giving  $\alpha \text{ explains } \delta \text{ bec\_poss } \{\alpha, \gamma_1\}$  (1a). The four optimal paths from  $\alpha$  to  $\delta$  are depicted, and non optimal paths (e.g. going through  $\beta_1$ ) exist also.

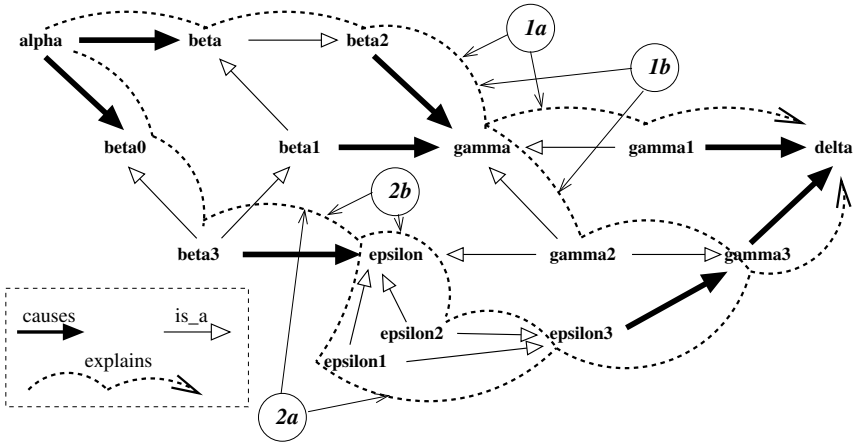


Fig. 1. Four optimal explanation paths from “alpha” to “delta”

### 3 An ASP Translation of the Formalism

#### 3.1 Presentation

We describe a program in DLV-Complex. A first version [8] used pure DLV, and was much slower and harder to read. We have successfully tested an example with more than a hundred symbols and more than 10 different explanation atoms for some  $(I, J)$  (made from two copies of the example of the diagram, linked through a few more data). We have encountered a problem, not listed in the “three problems” evoked above. The full program, including “optimization of the result”, did not work on our computer. The simplification step and the verification step are clearly separated from the first generating step, thus we have split the program in three parts: The first one generates various explanation atoms (including all the optimal ones). The second program keeps only the optimal explanation atoms, in order to help reading the set of the solutions, following Point 3c’. The third program checks whether the set of conditions is satisfied in the answer set considered. Then our “large example” works. Notice that it is useful to enumerate the answer sets (as described in the simple and very interesting [11]) in order to help this splitting.

#### 3.2 The Generating Part: Getting the Relevant Explanation Atoms

The “answer” of an ASP program is a set of *answer sets*, that is a set of concrete literals satisfying the rules (see e. g. [1] for exact definitions and [6] for the precise syntax of DLV (“:-” represents “ $\leftarrow$ ” and “-” alone represents “standard negation” while “ $\vee$ ” is the disjunction symbol in the head of the rules, and “ $\neq$ ” is “ $\neq$ ”). The user provides the following data:

- symbol(alpha) for each propositional symbol alpha.
- cause(alpha,beta) for each causal atom  $\alpha$  causes  $\beta$ ,

ont(alpha,beta) for each “is\_a” atom  $\alpha \rightarrow_{IS-A} \beta$  and

true(alpha) for each other propositional atom  $\alpha$  involved in formulas.

Causal and propositional formulas, such as  $(\neg\epsilon1 \wedge \neg\epsilon2) \vee \neg\gamma1 \vee \neg\gamma2$  must be put in conjunctive normal form, in order to be entered as sets of clauses:

```
{ -true(epsilon1) v -true(gamma1) v -true(gamma2). ,
  -true(epsilon2) v -true(gamma1) v -true(gamma2). }
```

The interesting result consists in the explanation predicates:

ecSet(alpha,beta,{alpha,delta,gamma}) represents the explanation atom  $\alpha$  explains  $\beta$  bec\_poss  $\{\alpha, \delta\}$ .

Here come the first rules (cf 2b §2.2):

```
ontt(I,J) :- ont(I,J). ont(I,K) :- ont(I,J), ont(J,K). ont(I,I) :- symbole(I).
```

We refer the reader to [9] for more details about the program. As an example of the advantage of using sets, let us give here the rule dealing with transitivity of explanations (cf 3b §2.2), ecinit referring to explanations not using transitivity rule. ( $\#insert(Set1,E2,Set)$  means:  $Set = Set1 \cup \{E2\}$ ):

```
ecSet(I,J,{I}) :- ecinit(I,J,I). ecSet(I,J,{I,E}) :- ecinit(I,J,E), not ecSet(I,J,{I}).
ecSet(I,J,Set) :- ecSet(I,K,Set1), not ecSet(I,J,Set1), ecinit(K,J,E2), E2 != K,
  #insert(Set1,E2,Set).
ecSet(I,J,Set) :- ecSet(I,K,Set), ecinit(K,J,K).
```

### 3.3 Optimizing the Explanation Atoms

The “weak simplification” rule 3c’ §2.2) is used at this step, omitted here for lack of space. Moreover, if two sets of condition  $\Phi, \Psi$  exist for some  $\alpha$  explains  $\beta$ , and if  $\Phi \models \Psi$  and not conversely, then only  $\alpha$  explains  $\beta$  bec\_poss  $\Psi$  is kept, the stronger set  $\Phi$  being discarded. This avoids clearly unnecessary explanation atoms. This part (not logically necessary) is costly, but helps interpreting the result by a human reader.

### 3.4 Checking the Set of Conditions

Finally, the following program starts from the result of any of the last two preceding programs and checks, in each answer set, whether the set of conditions is satisfied or not. The result is given by explVer(I,J,Set):  $I$  explains  $J$  bec\_poss  $Set$  where  $Set$  is satisfiable in the answer set considered (“Ver” stands for “verified”).

```
explSuppr(I,J,Set) :- ecSetRes(I,J,Set), -true(E), #member(E,Set).
```

```
explVer(I,J,Set) :- ecSetRes(I,J,Set), not explSuppr(I,J,Set).
```

Only “individual” checking is made here, in accordance with the requirement that the computational properties remain manageable.

With the whole chain (§3.2, 3.3, and 3.4), modifying a rule of the formalism can be done easily. The gain of using DLV-Complex instead of pure DLV (or gringo/claspD) is significant and worth mentioning.

### 3.5 Conclusion and Future Work

We have shown how the recent versions of running ASP systems allow easy translation of logical formalisms. The example of the explanation formalism

shows that such a translation can already be useful for testing new theories. In a near future, cases from the “real world” should be manageable and the end user should be able to use ASP for a great variety of diagnostic problems.

Here are two considerations about what could be hoped for future ASP systems in order to deal easily with this kind of problem. Since ASP systems are regularly evolving, we can hope that a near future the annoying trick consisting in launching the programs one after the other, and not in a single launch, should become unnecessary. It seems easy to detect that some predicates can safely be computed first, before launching the subsequent computation. In our example, computing `ecSet` first, then `ecSetRes` and finally `ecSetVer` is possible, and such one way dependencies could be detected. The great difference in practice between launching the three programs together, and launching them one after the other, shows that such improvement could have spectacular consequences.

Also, efficient “enumerating meta-predicates” would be useful (even if logically useless). A last interesting improvement would concern the possibility of implementing “enumerating answer sets”, as described in e.g. [11]. In this way real “brave” and “cautious reasoning could be envisioned, and much more.

For what concerns our own work, the important things to do are to apply the formalism to real situations, and, to this respect, firstly to significantly extend our notion of “ontology” towards a real one.

## References

1. Baral, C.: Knowledge representation, reasoning and declarative problem solving. Cambridge University Press, Cambridge (2003)
2. Besnard, P., Cordier, M.-O., Moinard, Y.: Ontology-based inference for causal explanation. *Integrated Computer-Aided Engineering J.* 15(4), 351–367 (2008)
3. Calimeri, F., Ianni, G.: Template programs for Disjunctive Logic Programming: An operational semantics. *AI Communications* 19(3), 193–206 (2006)
4. Giunchiglia, E., Lee, J., Lifschitz, V., McCain, N., Turner, H.: Nonmonotonic Causal Theories. *Artificial Intelligence* 153(1-2), 49–104 (2004)
5. Halpern, J., Pearl, J.: Causes and Explanations: A Structural-Model Approach - Part II: Explanations. In: *IJCAI 2001*, pp. 27–34. Morgan Kaufmann, San Francisco (2001)
6. Leone, N., Pfeifer, G., Faber, W., Eiter, T., Gottlob, G., Perri, S., Scarcello, F.: The DLV System for Knowledge Representation and Reasoning. *ACM Trans. on Computational Logic (TOCL)* 7(3), 499–562 (2006)
7. Mellor, D.H.: *The Facts of Causation*. Routledge, New York (1995)
8. Moinard, Y.: An Experience of Using ASP for Toy Examples. In: *Advances in Theory and Implementation ASP 2007*, pp. 133–147. Fac. de Ciencias, Univ. do Porto (2007)
9. Moinard, Y.: Using ASP with recent extensions for causal explanations. In: *AS-POCP Workshop, Associated With ICLP (2010)*
10. Shafer, G.: Causal Logic. In: Prade, H. (ed.) *ECAI 1998*, pp. 711–720 (1998)
11. Tari, L., Baral, C., Anwar, S.: A Language for Modular Answer Set Programming. In: *ASP 2005, CEUR Workshop Proc.*, vol. 142. CEUR-WS.org publ. (2005)



# Earthquake Prediction Based on Levenberg-Marquardt Algorithm Constrained Back-Propagation Neural Network Using DEMETER Data

Lingling Ma<sup>1</sup>, Fangzhou Xu<sup>1,2</sup>, Xinhong Wang<sup>1</sup>, and Lingli Tang<sup>1</sup>

<sup>1</sup> Academy of Opto-Electronics, Chinese Academy of Sciences,  
100190 Beijing, China  
llma@aoe.ac.cn

<sup>2</sup> Graduate University of China Academy of Sciences,  
100049 Beijing, China

**Abstract.** It is a popular problem that the mechanisms of earthquake are still not quite clear. The self-adaptive artificial neural network (ANN) method to combine contributions of various symptom factors of earthquake would be a feasible and useful tool. The back-propagation (BP) neural network can reflect the nonlinear relation between earthquake and various anomalies, therefore physical quantities measured by the DEMETER satellite including Electron density (Ne), Electron temperature (Te), ions temperature (Ti) and oxygen ion density (NO+), are collected to provide sample sets for a BP neural network. In order to improve the speed and the stability of BP neural network, the Levenberg-Marquardt algorithm is introduced to construct the model, and then model validation is performed on near 100 seismic events happened in 2008.

**Keywords:** Earthquake Prediction, Detection of Electro-Magnetic Emission Transmitted from Earthquake Regions (DEMETER), Back-propagation Neural Network, Levenberg-Marquardt Algorithm.

## 1 Introduction

Nowadays, how to accurately predict earthquake becomes an important problem that scientists are desired to resolve. However prediction of earthquake is up to now a great challenge especially in medium-term, short-term and impending earthquake prediction, and relevant studies are still in preliminary level which is far from practical applications. The difficulty of forecasting earthquake lies both in rough knowledge of mechanisms of earthquake preparation, and in the complexity of factors of earthquake precursors. The complicated correlation between earthquake precursory anomalies and earthquake occurrence usually behaves spatial heterogeneity, modal variety and regional diversity [1].

The neural network system is a highly adaptive nonlinear dynamic system. It does not require the object under analysis to satisfy a certain law, but by learning a large number of samples to extract the implied causality in the samples. Since 1990's, scientists have used neural network method in earthquake prediction in a series of

studies, mainly by using information of earthquake itself or observation data from ground stations as the predicting indicators. Conventional ground observation method can not continuously acquire multi-element precursory anomaly information in spatial and temporal, which leads to constrains on the construction of neural network model. The DEMETER (Detection of Electro-Magnetic Emission Transmitted from Earthquake Regions) satellite is the first solar synchronous satellite dedicated to monitoring seismic electromagnetic radiations. A large amount of observation data recorded since the launch of DEMETER can provide rich resources for the construction of neural network model [2].

In this paper, DEMETER data of various physical quantities acquired before seismic events are used to construct the sample sets, on which the back propagation neural network model is trained based on Levenberg-Marquardt algorithm, and then model validation is performed on those seismic events happened in 2008.

## 2 Theoretical Basis

### 2.1 Principles of Back-Propagation Neural Network

Back-propagation neural network is a common method of teaching artificial neural networks how to perform a given task. The back-propagation network necessarily has multilayer perceptions (usually with one input, one hidden, and one output layer), a general structure of which is shown in Fig. 1 [3].

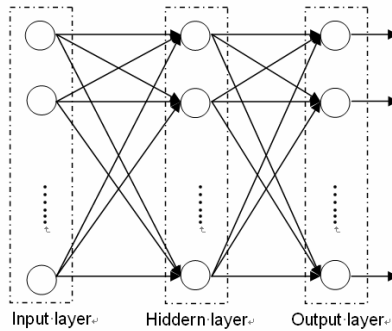


Fig. 1. General structure of a back-propagation network

The back-propagation learning is based on the gradient descent along the error surface. The Mean Absolute Error (MAE) can be used as a measure of the error made by the neural network.

$$E(W) = \frac{1}{2} \sum_{i=1}^N e_i^2(W) = \frac{1}{2} \sum_{i=1}^N (t_i - a_i)^2 \tag{1}$$

The above equation expresses the gap between desired output and practical output of each neuron. Here  $t_i$  is  $i$ th-neuron's objective value at the output layer, and  $a_i$  is  $i$ th-neuron's calculated value at the output layer, and  $N$  is the dimension of output vector.

### 2.2 Levenberg-Marquardt Algorithm

The Levenberg-Marquardt (L-M) learning algorithm provides a better learning means for the back-propagation network. It gives a good tradeoff between the speed of the Newton algorithm and the stability of the steepest descent method.

Assuming  $W(k)$  denotes the weight matrix at iterative turn  $k$ , the updated weight matrix  $W(k+1)$  in neural networks is calculated by L-M algorithm according to the two following basic rules [4],

$$W(k+1) = W(k) + \Delta W(k) \tag{2}$$

$$\Delta W = -[J^T(W)J(W) + \mu I]^{-1}J(W)e(W) \tag{3}$$

Where  $W = [w_1, w_2, \dots, w_N]$  consists of all weights of the network,  $e$  is the error vector comprising the errors for all the training examples,  $J$  is the Jacobian matrix,  $\mu$  is the learning rate which is to be updated using decay rate depending on the outcome, so as to control the behavior of the algorithm. The parameter  $\mu$  is a scalar controlling the behavior of the algorithm. When  $\mu = 0$ , the algorithm follows Newton’s method, using the approximate Hessian matrix. When  $\mu$  is high, the gradient descents with a small step size.

### 2.3 Calculation of Sample Data

Scientists acknowledged that a seismic electromagnetic anomaly usually is a climax of some process. These events are so sophisticated that the dynamic relations between their parameters result in high uncertainties in their prediction. Therefore, it is necessary to find more reliable methods to closely study the process and its relevant parameters.

Referring to Nemeč (2008)’s research, for a measured physical quantity  $E_i$ , its cumulative probability  $Fi$  is defined as the value of the corresponding cumulative distribution function, namely<sup>[5]</sup>,

$$Fi = \int_{-\infty}^{E_i} f(E)dE \tag{4}$$

Where the numerical range of  $Fi$  is  $[0,1]$ , which indicates the probability of occurrence of signals with an intensity less or equal to the measured level.

The distribution of the wave intensity in absence of seismic activity is called background. Background data itself was affected by the seasons, latitude and longitude, day and night, the solar activity and so on. Then the cumulative probability density function is calculated within each grid with the partition size of  $12 \times 12$  degree. The global background field is constructed according to different seasons: one is from October to April, the other is from May to September. Consequently the probabilistic intensity  $Ib$  of measure points over each earthquake place can be calculated with eq.5,

$$Ib = \frac{\sum_{i=1}^{M_b} Fi}{M_b} - 0.5 \tag{5}$$

where  $Mb$  is the number of cumulative probabilities  $Fi$  collected in a given grid. If the observed intensities were significantly lower or larger than the usual ones, the attributed cumulative probabilities would be significantly different from 0.5 and the resulting probabilistic intensity would be significantly different from 0.

### 3 A Case Study of Earthquake Prediction Based on L-M Algorithm Constrained Neural Network

#### 3.1 Selection of Sample Sets

Using IAP and ISL experiments onboard DEMETER satellite, the variations of the electron and ion densities have been statistically analyzed and disturbances were discovered in the vicinity of large earthquakes prior to the events. For example, the electron density measured by the ISL experiment at night detects anomalous variations significantly before the earthquakes [6].

In this study, the Electron density (Ne), Electron temperature (Te), ion temperature (Ti), and oxygen ion density (NO+), hydrogen ion density (H+), helium ion density (He+) from IAP and ISL experiments observed within 2007 and 2008 are used. During the analyzed period, the earthquakes with magnitude larger than or equal to 6.0 occurred all over the world are considered. For each seismic event, DEMETER data located within  $2^\circ \times 2^\circ$  (longitude by latitude) grid over the epicenter and acquired within the 30 days before the earthquake are considered.

Then, a detected vector  $v$  of DEMETER related with the occurrence of earthquake will be derived.

$$v = \langle Te, Ne, Ti, NO+, H+, He+ \rangle$$

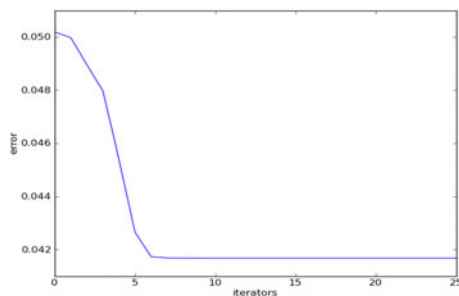
Here the original physical quantities are replaced with the  $Ib$  using eq.5 in which the background field has been cleared away, so as to remove the regular electromagnetic turbulence.

#### 3.2 Construction of Neural Network

The BP neural network are constructed using training sample sets in which the DEMETER data acquired in 2007. The input data consists of two parts of samples: one is data of earthquake with the magnitude higher than 6.0 and within 30 days before the earthquake, the other is data of areas not occurred earthquake.

The BP neural network used in this paper has three layers. Sigmoid function is employed in the hidden layer to generate a degree of non-linearity between the neuron's input and output. The number of units is determined as 20 by test. As far as output layer is concerned, the linear activation function is adopted and output unit is only one with the value of 0 or 1, representing of the magnitude less or higher than 6.0. The L-M algorithm is used in the training process, so the neural-network can converge fast in the training process.

The evolution of mean squared error during the training process is presented in the following figure, and the training time is 70.562 seconds.



**Fig. 2.** The evolution of mean squared error during the training process

We can see from the above figure that the training error descends quickly during the training process and becomes stable (less than 4.2%) after the sixth iteration, which indicates the high-speed convergence rate with L-M algorithm.

### 3.3 Validation of Earthquake Prediction

The test data is generated through the same method as section 3.3 but with the DEMETER data acquired in 2008. The test set consists of 191 vectors, 93 of which are from seismic area, 88 of which are from non-seismic area. The constructed neural network is validated over the test set, and the prediction results are compared with the seismic facts, as the following table shows:

**Table 1.** Prediction results of the constructed neural network

Prediction consequences	Earthquake	Non-earthquake
Earthquake	78	41
Non-earthquake	15	47
Accuracy	83.9%	46.6%

For 93 seismic area, 78 of which are predicted as “earthquake”, 15 of which are predicted as “non-earthquake”, the accuracy of prediction in the test case is 83.9%. For 88 non-seismic area, 41 of which are predicted as “earthquake”, 47 of which are predicted as “non-earthquake”, the accuracy of prediction in the test case is 46.6%. It proves a relatively good prediction with regard to earthquake area, while an undesirable prediction with regard to non-seismic area. The total accuracy is about 69%.

## 4 Conclusions

The attractive advantage of neural network method in the analysis of the complicated non-linear problems is its unnecessary demand of precise mathematic models and it could suit various environment factors through the adaptive network reconstruction,

which has been presented in the paper. However, neural network method is only a technical means and the effect on earthquake prediction depends on the chosen neural network model and the selected indicators. This paper tests a technical improvement of convergence rate and stability by introducing L-M algorithm according to the characteristics of complicated factors of earthquake events. While the indicators of the input and output of the network have good correlation, the forecasting accuracy is high, otherwise, the forecasting results could not be guaranteed even by the 100% verification of the study samples. Thus, in order to forecast the earthquake, the symptoms of the earthquake should be studied further, as well as the earth crustal physical processes. The artificial neural networks could get credible forecasting results only under the condition that the physical mechanism factors are well understood and the coupled elements can be separated from each other correctly.

## References

1. Jianmei, W., Wenzhong, Q.: BP Neural Network Classifier Based on Levenberg-Marquardt Algorithm. *Geomatics and Information Science of Wuhan University* 30(10), 928–931 (2005)
2. Parrot, M.: The micro-satellite Demeter: Data registration and data processing. *Seismo Electromagnetics: Litho-sphere-Atmosphere-Ionosphere Coupling*, 391–395 (2002)
3. Backpropagation Neural Network,  
<http://www.learnartificialneuralnetworks.com>
4. Suratgar, A.A., Tavakoli, M.B.: Modified Levenberg-Marquardt Method for Neural Networks Training. *World Academy of Science, Engineering and Technology* 6, 46–48 (2005)
5. Němec, F., Santolík, O., Parrot, M., Berthelier, J.J.: Spacecraft observations of electromagnetic perturbations connected with seismic activity. *Geophysical Research Letters* 35, L05109, 5 (2008)
6. Akhoondzadeh, M., Parrot, M., Saradjian, M.R.: Electron and ion density variations before strong earthquakes ( $M > 6.0$ ) using DEMETER and GPS data. *Natural Hazards and Earth System Sciences* 10, 7–18 (2010)

# Affinity Propagation on Identifying Communities in Social and Biological Networks

Caiyan Jia, Yawen Jiang, and Jian Yu

School of Computer and Information Technology, Beijing Jiaotong University  
Beijing 100044, P.R. China

{cyjia,09112077,jianyu}@bjtu.edu.cn

**Abstract.** Community structure is one of the most important features of complex networks, it uncovers the internal organization of the nodes. Affinity propagation (AP) is a recent proposed powerful cluster algorithm as it costs much less time and reaches much lower error. But it was shown that AP displayed severe convergence problems for identifying communities on the majority of unweighted protein-protein interaction (PPI) networks. On the contrary, AP was shown to achieve great success for identifying communities in benchmark artificial and social networks. So, in this study, we use AP to identify communities on artificial, social and unweighted PPI networks for finding the problem of the conflict. And we compare AP with Markov cluster (MCL), which was shown to outperform a number of clustering algorithms for PPI networks. The experimental results have shown that AP performs well without oscillations when similarity matrixes are chosen properly, and MCL is more accurate than AP but it runs slower than AP in large scale networks.

## 1 Introduction

Many complex networks in reality can be divided into communities or modules, where links within modules are much denser than those across modules, e.g. social networks, co-author networks, biological networks, etc. Identifying communities in complex networks is very important for real applications [1-2].

There are enormous methods to discover communities in networks [2]. We classify the current existing methods into two classes. One is the methods based on cluster analysis, such as k-means [3], hierarchical clustering [1], spectral clustering [4], AP clustering [5-6] and so on. First, similarities between pairs of nodes are defined. Then, clustering algorithms are used to clustering networks. The other is the methods designed for partitioning a network into some subgraphs by using the topology of the network. The most famous algorithms of this kind are GN [1], which continuously deletes edges with maximal betweenness, and CNM based on GN for maximizing modularity of a network [7].

With the research on community discovery, many methods have been developed and used in extracting modules from PPI networks [8]. And for making a comparative assessment, Brohée & Helden compared the four algorithms including Markov Cluster (MCL), Restricted Neighborhood Search Clustering

(RNSC), Super Paramagnetic Clustering (SPC), and Molecular Complex Detection (MCODE). The result showed MCL outperformed three other algorithms for identifying communities in PPI networks [9].

Recently, Frey & Dueck (2007) proposed a novel clustering algorithm, affinity propagation (AP), using similarity matrix [10]. The algorithm has been proved powerful as it costs much less time and reaches much lower error. Based on the result of Brohé and Helden, Valsblom & Wodak compared MCL with AP also for identifying communities in PPI networks [11]. The result showed that AP displayed severe convergence problems on the majority of unweighted PPI networks. So, Valsblom & Wodak concluded that 'AP as it stands, is not suitable for unweighted networks'. Meanwhile, Liu et al. [5] and Lai et al. [6] used AP to detect communities on some benchmark artificial and social networks, and achieved great success. It seems that the result conflicts with that of Valsblom & Wodak.

So, in this study, we compare MCL with AP for identifying communities on benchmark artificial networks, social networks and noisy versions of the unweighted PPI network with 408 *S. cerevisiae* protein complexes used in [11] for finding the problem of the conflict. And the results of GN with actual number of communities are used as the baseline since GN is a classical algorithm for partitioning social networks and the standard GN has been proved to be reliable to detect functional modules in PPI networks [12]. The experimental results have shown that AP performs well without oscillations when similarity matrixes are chosen properly, and MCL is more accurate than AP but AP often runs several times faster than MCL on large scale networks. And we have found that AP failed to extract modules in PPI networks in the previous study because the similarity matrixes were just the adjacency matrixes, which didn't consider any topological structure of the whole networks while clustering methods are always sensitive to similarity metric for pairs of data.

The rest of paper is organized as follows. Section 2 introduces some preliminaries related to community discovery and gives a brief description for AP and MCL clustering algorithms. Section 3 shows some experimental results. Section 4 concludes the paper.

## 2 Community Structure and the Clustering Algorithms

Let  $G = \{V, E\}$  be a network, where  $V$  is a set of  $n$  vertices and  $E$  is a set of  $m$  edges. The network can be represented by an adjacency matrix  $A$  with  $A(i, j) = 1$  for  $(i, j) \in E$ , otherwise  $A(i, j) = 0$ . A *community* is a subgraph of a network whose nodes are more tightly connected with each other than with nodes outside the subgraph.

Quantifying similarity between vertices in a network is a key problem for clustering the network into groups. *Structural equivalence* is a commonly used approach. Two vertices are considered *structurally equivalent* if they share many of the same network neighbors. There are many ways to define similarity based on *structural equivalence* such as *Jaccard* index, *cosine* similarity and so on [13].



But the similarity metrics only use information of directly connected neighbors of pairs of vertices. It is possible that a vertex  $j$  is similar to vertex  $i$  if  $i$  has a neighbor  $v$  that is itself similar to  $j$ . So *regular equivalence* was proposed to use the whole topological structure of a network. Based on *regular equivalence*, an iterative method was given in [13] to obtain the similarity matrix of a network. The similarity is in fact a weighted count of the number of paths of all lengths between the vertices, is selected as the similarity metric for AP in the study.

AP is a recent proposed clustering algorithm. In the algorithm, every cluster can be represented by one of its nodes in this cluster, which is called its exemplar. During the clustering process of AP, every node in the network can be considered as a candidate exemplar, which iteratively competes with other candidate exemplars by maximizing the sum of the responsibility  $r(i, k)$  and the availability  $a(i, k)$  of the node, where

$$\begin{aligned} r(i, k) &= s(i, k) - \max_{j \neq k} \{a(i, j) + s(i, j)\}, \forall i, k, \\ a(i, k) &= \min\{0, r(k, k) + \sum_{j \notin \{i, k\}} \max\{0, r(j, k)\}\}, i \neq k, \\ a(k, k) &= \sum_{j \neq k} \max\{0, r(j, k)\}, \end{aligned}$$

$s(i, k)$  is the similarity of nodes  $i$  and  $k$  for  $i \neq k$  and  $s(k, k) = p$ ,  $p$  is a preference value and determines the number of clusters. However, directly maximizing the sum of  $r(i, k)$  and  $a(i, k)$  could lead to numerical oscillations in some cases. Damping factor  $\lambda$  was introduced to alleviate oscillations (see [10] for details).

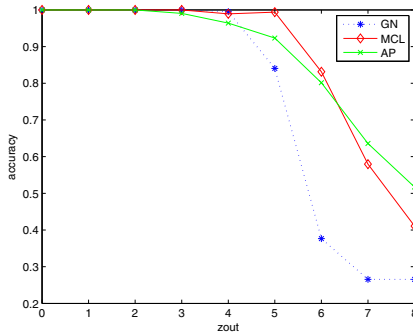
Markov cluster (MCL) algorithm was designed for graph clustering, the same as community discovery for a networks [14]. It was proposed to simulate random walks on an underlying interaction network by alternating two operations, expansion and inflation, based on the observation that a random walk in a network that visits a dense cluster will likely not leave the cluster until many of its vertices have been visited. When a graph is translated into a stochastic Markov matrix, expansion operation considers longer random walks one by one, and the inflation operation is used to weaken weak edges and strengthen strong ones. The operations are iteratively used until the algorithm converges to a steady state. MCL is a very simple algorithm. It just needs to input the adjacency matrix  $A$  of a network and tune inflation power to obtain the proper number of clusters.

### 3 Experimental Results

In this section, we will compare AP with MCL for community discovery in computer-generated networks, three popular real world social networks including the network of Zachary's karate club, the network of American college football team [1] and the network of bottlenose dolphins [15]. Furthermore, we will compare AP with MCL for identifying communities in the noisy versions of the unweighted PPI network built from 408 *S. cerevisiae* protein complexes hand curated in-house [16]. And in all of our experiments, the results of GN with actual number of communities are used as the baseline. All experiments are performed on an Intel computer with 2 GHz processor and 2GB main memory. The operating system is Windows XP.

### 3.1 Computer Generated Networks

Following Grivan and Newman [1], we generate a set of artificial networks with 128 vertices divided into 4 communities. The average degree of each vertex is 16 and the average number of edges of each vertex between communities, denoted by  $z_{out}$ , is varied from 0 to 8. For each  $z_{out}$ , five random networks are feeded to AP, MCL and GN. We compare the average accuracy of the networks, the fraction of vertices that are classified into their correct communities [1,5-6], for the three algorithms on  $z_{out}$  from 0 to 8. The results are shown in Fig. 1. In all of the experiments, convergence factor  $\alpha$  is set to be 0.9 for getting the similarity matrixes of AP by means of the iterative method in [13].



**Fig. 1.** Performance for identifying communities in artificial networks

Fig. 1 shows both of MCL and AP are more accurate than GN especially when  $z_{out}$  is large. And MCL is slightly better than AP when  $z_{out} \leq 6$ , but is worse than AP when  $z_{out} > 6$ . So AP might have stronger ability to discriminate weak communities than MCL.

### 3.2 Real Social Networks

The results of AP, MCL and GN on the network of Zachary’s karate club, the network of American college football team and the network of bottlenose dolphin are shown in table 1, where all the best results are shown in bold.

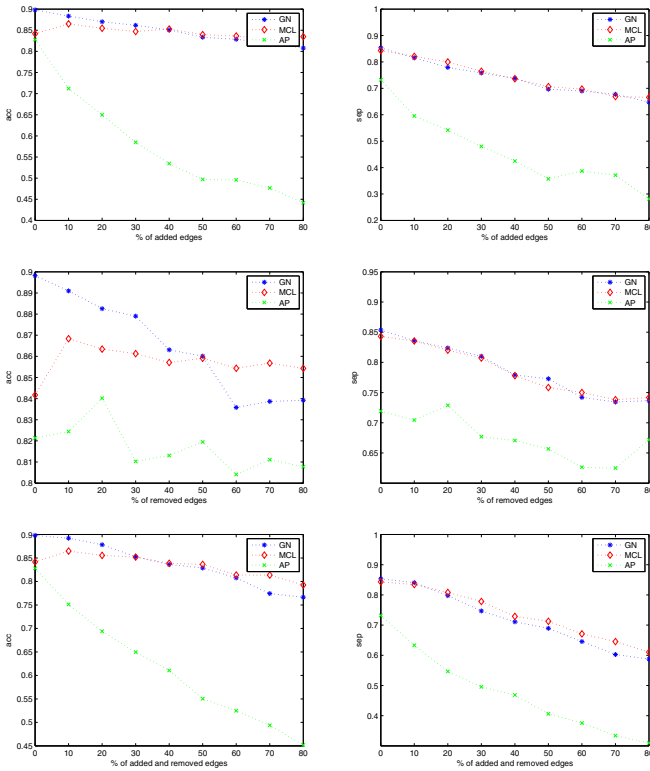
**Table 1.** Performance for identifying communities in social networks

	GN	AP	MCL
Karate	0.9706	<b>1.0</b>	0.9706
Football	0.9043	0.9043	<b>0.9130</b>
Dolphin	<b>0.9839</b>	0.9677	0.9677

In table 1 we show both of MCL and AP have satisfactory results. GN achieves the best accuracy on the dolphin network since the golden standard comes from the partition of GN in company with expert knowledge [15].

### 3.3 Protein-Protein Interaction Networks

Also, we choose the PPI network built from 408 *S. cerevisiae* protein complexes as our testing set as the same as [11]. And we randomly add or remove a portion of edges to the original network to simulate noise, use *acc* and *sep* to evaluate the performance of the algorithms after singletons are removed [11]. The comparison results with different noise ratios are shown in Fig. 2, where  $x\%$  of added and removed edges means we add and remove half of  $x\%$  edges, respectively. And the damping factor  $\lambda$  of AP is defaulted as 0.9 in all of the experiments [10].



**Fig. 2.** Robustness of the algorithms to random edge addition and removal

Fig.2 shows AP performs well without oscillations at all noise ratios. But why did AP oscillate in [11]? The input pairwise 'similarities' were defined as  $s(i, j) = 1$  if protein  $i$  and  $j$  were annotated to the same complex. In other words, similarity matrixes were just the adjacency matrixes of the networks. They didn't consider the topology of the networks. And GN and MCL both perform better than AP. But GN needs to specify the number of clusters and runs very slowly. For an instance, it runs more than 18 hours when we add 20%

edges to the network. And MCL often runs several times slower than AP. So AP is more suitable for partitioning large scale networks than MCL.

## 4 Conclusion

In the study, we compare MCL with AP for identifying communities on artificial, social and PPI networks. The experimental results have shown AP performs well without oscillations and MCL is more accurate but runs slower than AP.

**Acknowledgements.** This work was supported in part by NSFC (Grant No. 60875031, 60905029 and 90820013), 973 Project (Grant No. 2007CB311002).

## References

1. Girvan, M., Newman, M.J.: Community structure in social and biological networks. *Proc. Nat. Acad. Sci.* 99, 7821–7826 (2002)
2. Fortunato, S.: Community detection in graphs. *Physics Reports* 486, 75–174 (2010)
3. Rattigan, M., Maier, M., Jensen, D.: Graph clustering with network structure indices. In: *Proc. of ICML 2007*, Corvallis, Oregon, pp. 783–790 (2007)
4. Jiang, J.Q., Dressb, A.W.M., Yanga, G.: A spectral clustering-based framework for detecting community structures in complex networks. *Applied Mathematics Letters* 22(9), 1479–1482 (2009)
5. Liu, Z.Y., Li, P., Zheng, Y.B., Sun, M.S.: Community detection by affinity propagation. Technical Report (2008)
6. Lai, D., Lu, H.T.: Identification of community structure in complex networks using affinity propagation clustering method. *Modern Physics Letters B* 22(16), 1547–1566 (2008)
7. Clauset, A., Newman, M.E.J., Moore, C.: Finding community structure in very large networks. *Phys. Rev. E* 70, 066111 (2004)
8. Li, X., Wu, M., Kwok, C., Ng, S.: Computational approaches for detecting protein complexes from protein interaction networks: a survey. *BMC Bioinformatics* 11, S3 (2010)
9. Brohee, S., Helden, J.V.: Evaluation of clustering algorithms for protein-protein interaction networks. *BMC Bioinformatics* 7, 488 (2006)
10. Frey, B.J., Dueck, D.: Clustering by passing messages between data points. *Science* 315, 972–976 (2007)
11. Vlasblom, J., Wodak, S.J.: Markov clustering versus affinity propagation for the partitioning of protein interaction graphs. *BMC Bioinformatics* 10, 99 (2009)
12. Dunn, R., Dudbridge, F., Sanderson, C.M.: The use of edge-betweenness clustering to investigate biological function in protein interaction networks. *BMC Bioinformatics* 6, 39 (2005)
13. Leicht, E.A., Petter, H., Newman, M.E.J.: Vertex similarity in networks. *Physical Review E* 73(2), 26120 (2006)
14. Dongen, S.V.: Graph clustering by flow simulation. PhD Thesis Centers for Mathematics and Computer Science (CWI), University of Utrecht (2000)
15. Lusseau, D., Newman, M.E.J.: Identifying the role that animals play in their social networks. *Proc. R. Soc. Lond. B (suppl.)* 271, S477–S481 (2004)
16. Pu, S., Wong, J., Turner, B., Cho, E., Wodak, S.J.: Up-to-date catalogues of yeast protein complexes. *Nucleic Acids Res* 37(3), 825–831 (2009)

# Semantic Decomposition of Indicators and Corresponding Measurement Units

Michaela Denk<sup>1</sup> and Wilfried Grossmann<sup>2</sup>

International Monetary Fund, Statistics Department\*,  
1900 Pennsylvania Ave NW, Washington, DC, 20431, USA  
mdenk@imf.org

University of Vienna, Institute for Scientific Computing,  
Universitaetsstr. 5, 1010 Wien, Austria  
wilfried.grossmann@univie.ac.at

**Abstract.** Based on a review of existing standards and guidelines as well as the current international practice of modeling measurement units and related concepts in representation of economic and statistical data, the broad notion of unit of measure is decomposed into its basic building blocks and their interrelations. As these components are often presented as part of the measured indicator, the indicators themselves were included in the analysis. The resulting semantic model is regarded as a contribution to the further development of the content oriented guidelines in terms of harmonizing cross-domain concepts such as the unit of measure.

**Keywords:** Semantic decomposition, data exchange, knowledge management, content oriented guidelines, knowledge dissemination.

## 1 Introduction

One of the basic procedures in knowledge management is exchange of information. In this note we will consider some issues in exchange of numerical information between different systems. An important fact in this case is proper understanding of the exchanged numbers, more precisely what is represented by the number and what units have been used for measurement. At a first glance this problem seems to be rather easy and already resolved. Standardization of units of measure from a primarily scientific perspective is the aim of the International System of Units [11] and the Unified Code for Units of Measure [12]. The basic idea of these systems is starting with a number of **base** units, like length mass time, and obtaining from such base units so called **derived** units defined by formal relations between measured (base) quantities. Furthermore “**dimensionless**” units arise from a ratio of two quantities measured in the same unit.

In connection with business applications a number of additional units like currency or index numbers for economic performance occur besides the units considered in the

---

\* The views expressed herein are those of the authors and should not be attributed to the IMF, its Executive Board, or its management.

above mentioned systems. Due to the fact that such information is of utmost importance in analyzing global economic problems it is not surprising that organizations like IMF, OECD, UN, or World Bank are interested in defining standards for exchanging such measurements. An attempt to standardize this information resulted in the so called Statistical Data and Metadata Exchange (SDMX) [13]. However, a look at the given specification and existing practice at different organizations shows that application of the guidelines is by no means standardized. A major problem seems to be that besides definition of the unit of measurement in a narrower sense this unit of measurement is used as a vehicle for transporting some additional information about the figure, which is important for proper understanding.

This paper makes an attempt to the semantic decomposition of the unit of measurement for describing numerical information. Following the terminology used in economics we will call these numerical information **indicators**. In section 2, we briefly summarize some empirical facts about usage based on a review of 14 different organizations and in section 3 we specify a formal model which allows precise definition of all investigated examples. Section 4 gives an outlook on potential future developments.

## 2 Existing Guidelines and Current Practice

The SDMX COG [13] can be regarded as the most prominent current effort focusing on the harmonization of cross-domain concepts for data exchange. They recommend practices for creating interoperable data and metadata sets using the SDMX technical standards with the intent of generic applicability across subject-matter domains. Four cross-domain concepts described in the SDMX COG are related to the unit of measure, viz. unit of measure (item 65), adjustment (3), base period (5), and unit multiplier (64).

The representation of unit of measure, adjustment, and base period is divided into a code and free text, but no code lists are provided. For base period, a time stamp is also required. Unit multiplier is the only related concept with a code list available. **Unit of measure** is defined by three components, (i) type (currency is named as an example), (ii) unit code, and (iii) unit of measure detail. Type is not explained or used any further in the COG. For units that are index numbers, index type exists as a separate cross-domain concept. For unit itself, several examples are listed, e.g. kg, Euro, counts, and index numbers. The unit multiplier is referred to as a supplementary concept necessary to interpret observation values. However, no connection is established to other related concepts such as adjustment and base period. Rather, unit of measure detail is supposed to hold any additional information required to specify the unit in detail. **Unit multiplier** specifies the exponent to the basis 10 by which observation values were divided, usually for presentation purposes. For instance, a unit multiplier of 3 indicates that the data are provided in thousands. The code list does not contain negative unit multipliers. **Base period** is relevant for the interpretation of index data, series at real terms, e.g. data at constant prices, change measures such as percentage changes with respect to the previous period, or other series based on a certain point in time (in addition to the reference period). The base period of an index is the period when the index equals the base value (per definition). **Adjustment** is a concept that is included in the unit of measure or the economic indicator in many statistical databases, as illustrated in [2].

Despite the existence of standards and guidelines for representing the unit of measure, their adoption and implementation can only be observed to an unsatisfactory extent. Examples from the following databases of international organizations show that current practice is very diverse and would rather suggest a lack of such standards: Principal Global Indicators Website of the Inter-Agency Group on Economic and Financial Data [8], World Economic Outlook (WEO) of the International Monetary Fund [9], OECD.Stat Data Warehouse of the Organization for Economic Cooperation and Development (OECD), three datasets [10], Statistics Website of the Bank for International Settlements (BIS) [1], Statistical Data Warehouse of the European Central Bank (ECB) [3], Eurostat's Selected Principal European Economic Indicators (PEEI) [5], Eurostat's EUROIND Database [4], World Development Indicators Online Database (World Bank) [15], Millennium Development Goals Indicators Database (UNStats) [14], and three datasets from IMF Internal Databases.

From these 14 investigated datasets, four do not separate the economic indicator from the unit of measure or provide the unit information, whereas four other datasets even split other concepts such as unit multiplier or adjustment method from the unit. The other six databases separate unit of measure from economic indicator. A broad variety of unit types is used such as index, count, ratio, rate, percentage, or changes. Refer to the appendix of [2] for a list of all examples examined. The cases with a single, mixed dimension at least combine information on measured (economic) indicator, type of unit, unit of measure, adjustment method, and frequency. Several examples (e.g. "Personal computers" or "Youth unemployment rate, aged 15-24, men") even omit the unit information completely, assuming that it is obvious from the indicator used. On the other hand, observe that these "measurement units" give information about the underlying population to which the concept refers.

### 3 Semantic Decomposition and Metadata Model

Despite the imperfect compliance to the SDMX COG in dissemination practice, these guidelines still do not seem sufficiently specific and restrictive to account for the requirements derived from the investigated examples, especially concerning the decomposition of unit of measure and related concepts as well as the guidance on value domains (code lists) for the resulting components. Based on the examples, different types of units that follow a similar structure and can be broken down into the same components as well as basic, extendable value domains for these components as a foundation for the specification of suitable code lists are derived.

In order to define a unified structure we propose a generic model for semantic decomposition of the measure of units as it is shown in Figure 1. The proposal is based on the decomposition of the indicator and unit measure into four components, which can be applied recursively for describing complex indicators and corresponding measurement units:

- **Indicator** with components Type, Concept, and Population
- **Measurement** with components Type, Unit, Unit Multiplier, Statistical Measure and relations to Reference and Adjustment
- **Reference** with components Period, Value, and Statistical Measure
- **Adjustment** with components Price, Econometric, and Exchange Rate

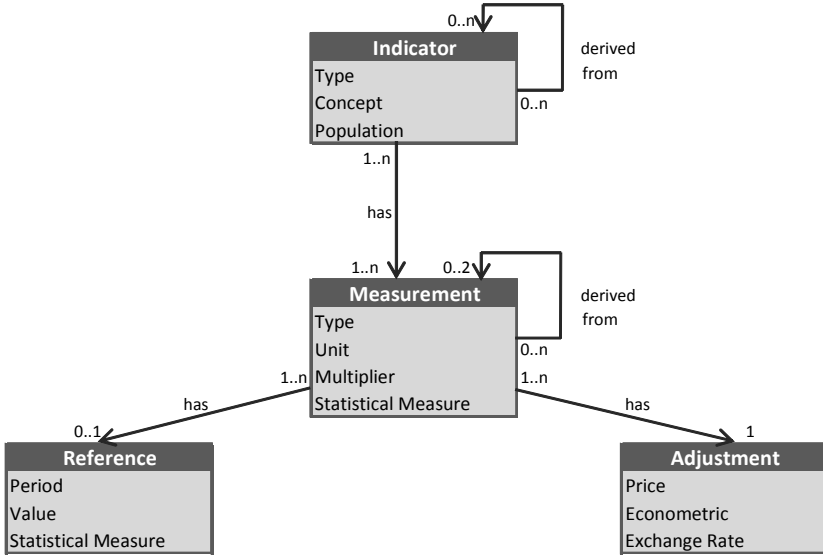


Fig. 1. Semantic Model for Indicators and Corresponding Measurement

Strictly speaking the **Indicator** itself is not part of the measurement and should not be included into the measurement. It corresponds more to the heading of the data which are exchanged or to the name of the attribute for which data are provided. The mentioned attributes **Type**, **Concept** and **Population** represent in some sense a minimal information for keeping the semantic of the data. For the **Type**-attribute two values are possible:

- Type = base: Base types indicators correspond to quantities which are available for direct measurement. In context of economic applications the most important quantities are measurements like count or currencies, but also base quantities of SI and/or UCUM may occur.
- Type = derived: The value derived is applicable to all indicators obtained from base types by application of arithmetic operations like ratio, product, difference, and sum. The derived quantities of SI/UCUM such as area, volume, and velocity can be regarded as subtypes thereof. Similarly, subtypes especially relevant in economic applications such as index, currency exchange rate, or interest rate, are introduced.

The **Concept** attribute states the definition of the attribute of interest and the **Population** attribute gives reference for which population the indicator may be applied.

Description of the **Measurement** is based on four attributes. The **Type**-attribute is similar to the type attribute of the indicator itself but has to take into account a more detailed specification. In case of derived type a detailed specification of the formula is of interest or usage of reserved keywords like index or fraction. The **Unit** attribute gives us specification of the measurement scale. For example in case of currencies it uses specific currencies like EUROS or USD. For derived indicators the unit can be



described by the applied formula leading to similar constructions as in case of the SI/UCUM specification. In some cases specific names for the derived units are in use, for example index points in case of differences between two indices. Usage of the **Multiplier** attribute was already described in section 2 and corresponds to traditional usage in economic data representation. The **Statistical Measure** attribute is of utmost importance in case of high frequency data, where specific summary measures like means, sums, or highest and lowest values of stock indices are exchanged.

The **Reference** component is mandatory for some measurement type like index or balance indicator, or any type of adjustment=constant. The value domain of (reference or base) **Period** includes time stamps (in different granularity) and predefined values such as previous period, corresponding period of previous year, years since time stamp, or statistical measure of years since time stamp. The **Value** attribute is used for specification of base values for the reference period. **Statistical Measure** is similar to the corresponding attribute in the measurement unit and most often an average.

The rationale behind the **Adjustment** is to capture an essential part of computation which is of interest for proper understanding of the figures. For Adjustment three possible adjustments are defined which cover the cases which have been investigated but the list can be extended according to specific needs.

Recursive application of the model in figure 1 allows detailed specification of measurement units for all types of all derived indicators. In such case references to underlying Measurement and/or Indicators are required for all derived types. In addition, **unit families**, **standard units**, and **conversion factors** [6], [7] play an important role, especially for physical (and currency) units. These concepts are omitted here due to the focus of the paper and the lack of space.

## 4 Conclusion

The motivation of this paper was the decomposition of the concept “unit of measure” into its basic building blocks, serving as a foundation for the development of standardized value domains and code lists for these sub-concepts. To this end, we carried out an analysis of existing standards and guidelines as well as the current practice of modeling unit of measure and related concepts, based on examples from databases of international organizations that act as sponsors of the SDMX initiative. From this investigation, we derived an extendable set of generic unit types and subtypes, assigned units, and a set of cross-domain concepts related to unit of measure and the measured indicator. These findings provide a sound basis for an extension and further development of the SDMX COG with respect to cross-domain concepts as well as code lists in the context of unit of measure. Future work will involve the extension of the presented ideas to elaborate in more detail the representation of derived indicators, measurement, and unit families as well as the required constraints and rules for the propagation of indicator and measurement information. The definition of code lists including “shortcut” descriptors for mixed dimensions/concepts is also of high priority. The development of a unit (type) calculus to better account for and make use of the calculability characteristic of units based on ideas developed by Froeschl [6] is the long-term objective of this research.

**Acknowledgments.** Special thanks go to our colleagues at the IMF Statistics Department, especially Gareth McGuinness, for valuable discussions and feedback.

## References

1. Bank for International Settlements Locational Banking Statistics,  
<http://www.bis.org/>
2. Denk, M., Grossmann, W., Froeschl, K.A.: Towards a best practice of modeling unit of measure and related statistical metadata. In: European Conference on Quality in Official Statistics. Statistics Finland, Helsinki (2010) (forthcoming),  
<http://q2010.stat.fi/papers/>
3. European Central Bank Data Warehouse, <http://sdw.ecb.europa.eu/>
4. Eurostat EUROIND Database,  
<http://epp.eurostat.ec.europa.eu/portal/page/portal/euroindicators/database>
5. Eurostat Selected Principal European Economic Indicators,  
<http://ec.europa.eu/eurostat/euroindicators>
6. Froeschl, K.A.: Metadata Management in Statistical Information Processing. Springer, New York (1997)
7. Froeschl, K.A., Grossmann, W., Del Vecchio, V.: The Concept of Statistical Metadata. METANET Deliverable D5, 127 (2003)
8. Inter-Agency Group on Economic and Financial Statistics Principal Global Indicators,  
<http://www.principalglobalindicators.org/>
9. International Monetary Fund World Economic Outlook Database,  
<http://www.imf.org/external/pubs/ft/weo/2009/02/weodata/index.aspx>
10. OECD, OECD Stat Extracts (2010), <http://stats.oecd.org/>
11. The International System of Units, <http://www.bipm.org/en/si/>
12. The Unified Code for Units of Measure, <http://unitsofmeasure.org/>
13. SDMX Content Oriented Guidelines, 16pp + 5 annexes (2009),  
<http://www.sdmx.org/>
14. UNStats Millennium Development Goals Indicators Database,  
<http://mdgs.un.org/>
15. World Bank World Development Indicators Online Database,  
<http://ddp-ext.worldbank.org/ext/DDPQQ/member.do?method=getMembers&userid=1&queryId=135>

# Engineering Knowledge for Assistive Living

Liming Chen and Chris Nugent

School of Computing and Mathematics  
University of Ulster, United Kingdom  
{l.chen, cd.nugent}@ulster.ac.uk

**Abstract.** This paper introduces a knowledge based approach to assistive living in smart homes. It proposes a system architecture that makes use of knowledge in the lifecycle of assistive living. The paper describes ontology based knowledge engineering practices and discusses mechanisms for exploiting knowledge for activity recognition and assistance. It presents system implementation and experiments, and discusses initial results.

**Keywords:** Smart home, ontology, knowledge engineering, activity recognition, assistive living.

## 1 Introduction

Smart Home (SH) [1] has emerged as a mainstream approach to providing assistive living and supporting ageing-in-place. A SH is considered to be augmented living environments equipped with sensors and actuators, within which monitoring of Activities of Daily Living (ADL) and personalised assistance can be facilitated. Though a number of Lab-based or real living SHs has been developed and an abundance of supportive technologies provide fragments of the necessary functionality [2], existing SH technologies and solutions suffer from major drawbacks, including data heterogeneity, lack of interoperability, reusability and applicability of technologies and solutions as well.

To address these problems, this paper introduces a knowledge based approach to evolving current smart home technologies towards the future infrastructure that is needed to support the application and large-scale deployment of smart homes in real world context. The approach is motivated by the observations that ADLs as daily routines are full of commonsense knowledge and heuristics providing rich links between environments, events and activities. The proposed approach aims to exploit semantic technologies to engineer SH domain knowledge. Specifically, SH resources, i.e., sensors, sensor data, actuators, inhabitants, ADL and services, will be formally modelled and explicitly represented with well-defined meaning, rich contextual and/or heuristic knowledge. As such, the approach can support resource interoperability and reusability through semantic descriptions, realise advanced features in the lifecycle of assistive living by making extensive use of semantic/knowledge-based intelligent processing techniques.

The paper is organised as follows. Section 2 introduces a knowledge based system architecture. Section 3 describes knowledge engineering and management practices.

Section 4 outlines some typical knowledge use scenarios. We present system implementation and experiments in Section 5 and conclude the paper in Section 6.

## 2 A Knowledge Enabled Approach to Assistive Living

Fig. 1 shows the proposed system architecture for a SH. The Physical Layer consists of physical hardware such as sensors, actuators, and various devices including medical equipment, household appliances and network components. This layer provides the means to monitor and capture the events and actions in a SH. The Data Layer collects and stores raw data in a number of data stores. These stores are usually disparate in data formats and access interfaces, with each of them being dedicated to individual application scenarios. The Application Layer contains application dependent services and systems for assistive living. Within this layer applications can process sensor data from the Data Layer and control actuators and/or devices in the Physical Layer to offer assistance. These three layers have so far been the major components underpinning existing SH application design and development. While each layer is indispensable for any SH application, the close coupling among sensors, data and applications, often having one to one, ad hoc relationships.

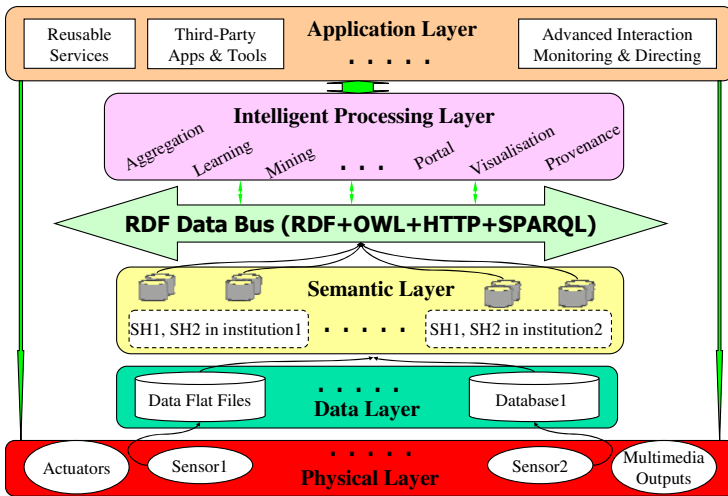


Fig. 1. The conceptual system architecture

The proposed approach incorporates a Semantic Layer, a RDF<sup>1</sup> Data Bus and an Intelligent Service Layer into the systems architecture. The goal of the Semantic Layer is to provide a homogeneous view over heterogeneous data, thus enabling seamless data access, sharing, integration and fusion across multiple organisations, providing interoperability and machine understandability. It achieves this by using SH ontologies as a unified conceptual backbone for data modeling and representation. Semantic modelling allows the markup of various data with rich metadata and semantics to generate

<sup>1</sup> RDF, OWL, HTTP, SPARQL are W3C standards, refer to W3C web site - [www.w3.org](http://www.w3.org)

semantic content. Multiple SHs in geographically distributed locations supported by various organisations can then aggregate and fuse their SH data. The uniform data models and representation, e.g., RDF or OWL, allow seamless data access through the RDF Bus based on the standard communication protocol HTTP and RDF query language SPARQL. The Semantic Layer is also responsible for providing tools and APIs for semantic data retrieval and reasoning.

The Intelligent Service Layer is built upon the semantic content and functionalities of the Semantic Layer. Its purpose is to exploit semantics and descriptive knowledge to provide advanced processing and presentation capabilities and services. The former provides added-values to the query interfaces of the RDF Bus through further analysis and reasoning over recorded SH data. The latter essentially visualises the contents of the repositories and the outputs of the processing services. The selection and use of such services will depend on the nature and availability of collected data as well as the personal needs of inhabitants and care providers, hence allowing for personalisation. They are accessible to third party developers, thus interoperable and reusable at both the service and application level.

### 3 Ontology-Based Knowledge Engineering and Management

A SH is a home setting where ADLs are usually performed in specific circumstances, i.e., in specific environments with specific objects used for specific purposes. For example, brushing teeth usually takes place two times a day, in a bathroom, normally in the morning and before going to bed. This activity usually involves the use of toothpaste and a toothbrush. As humans have different life styles, habits or abilities, individuals' ADLs and the way they perform them may vary one from another. Even for the same type of activity, e.g., making white coffee, different people may use different ingredients, and in different orders, e.g., adding milk first and then sugar, or vice versa. As such ADLs can be categorized as *generic* ADLs applicable to all and *personalised* ADLs with subtlety of individuals. In addition, ADLs can be conceptualized at different levels of granularity. For example, Grooming can be considered to be comprised of sub-activities *Washing*, *Brushing* and *Applying Make-up*. There are usually a "*is-a*" and "*part-of*" relationships between a primitive and composite ADL. All these observations can be viewed as prior domain knowledge and heuristics that can facilitate assistive living. The key is how to formally capture, encode and represent such domain knowledge.

We carry out knowledge acquisition through interviews, questionnaires and by studying existing documents from which we derive the conceptual models for describing activities and their relations with sensors and objects. Based on SH characterization and the conceptual activity model we develop ADL ontologies using Protégé [3] as shown in Fig. 2. The ADL ontology consists of an activity hierarchy in which each node, also called a class, denotes a type of ADL. Each class is described with a number of properties. In a similar way we develop SH context ontologies that consist of classes and properties for describing SH entities such as *Device*, *Furniture*, *Location*, *Time* and *Sensor*, and their interrelationships with an activity class. Each sensor monitors and reflects one facet of a situation. By aggregating individual sensor observations the contextual snapshots at specific time points, or say a situation, can be generated, which can be used to perform activity recognition.

Given the nature of sensor data in SH we develop a two phase semi-automatic approach to generating semantic descriptions. In the first phase data sources such as sensors and devices are manually semantically described. In the second phase dynamically collected sensor data are first converted to textual descriptors. They are then automatically attached to semantic instances of the corresponding ontological classes to create a semantic knowledge repository. All these operations are performed through demon-like style software tools embedded in the implemented system. the generated semantic data and metadata are archived in a knowledge repository.

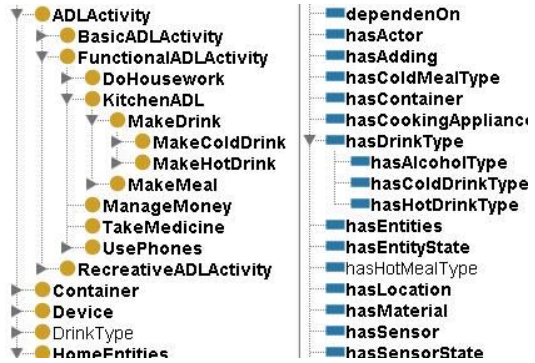


Fig. 2. A fragment of the ADL ontologies

### 4 Using Knowledge for Assistive Living

Once knowledge is modeled, captured and stored in knowledge repositories, it can be exploited in a diversity of ways. Three key use cases in the context of assistive living are described below.

**Activity Recognition** - In ontological SH modeling, activities are modeled as activity classes in the ADL ontologies and contextual information such as time, location and the entities involved is modeled as properties for describing activity classes. As such, a situation at a specific time point is actually a concept description created from SH contextual ontologies, denoting an unknown activity. In this case, activity recognition can be mapped to the classification of the unknown activity into the right position of the class hierarchy of the activity ontologies and the identification of the equivalent activity class. This can be mapped to the subsumption problem in Description Logic, i.e., to decide if a concept description  $\mathcal{C}$  is subsumed by a concept description  $\mathcal{D}$ , denoted as  $\mathcal{C} \sqsubseteq \mathcal{D}$ .

**Activity Model Learning** - As activity models play a critical role in mining real-time sensor data for activity recognition, complete and accurate activity models are of paramount importance. While ADL ontologies have the advantage of providing knowledge-rich activity models, it is difficult to manually build comprehensive ADL ontologies. In particular, given the complexity of ADLs, the differences of ways and capabilities of users carrying out ADLs and also the levels of granularity that an ADL can be modeled, building complete one-for-all ADL ontologies is not only infeasible but also inflexible for adapting to various evolving use scenarios. To address this problem, we can use the manually developed ADL ontologies as the seed ADL models. The seed activity models are, on one hand, used to recognize activities as

described above. On the other hand, we developed learning algorithms that can learn activity models from sensor activations and the classified activity traces. As such, ADL ontologies can grow naturally as it is used for activity recognition. This is actually a self-learning process in order to adapt to user ADL styles and use scenarios.

**Activity Assistance** - With activity ontologies as activity models, and activity instances from a specific inhabitant as the inhabitant's activity profile, the propose approach can support both coarse-grained and fine-grained activity assistance. The former is directly based on subsumption reasoning at concept (or class) level, while the latter on subsumption reasoning at instance level, i.e., based on an inhabitant's ADL profile. For coarse-grained activity assistance, the process is nearly the same as activity recognition. The extra step is to compare the properties of the recognized activity with the properties identified by sensor observations. The missing property(ies) can then be used to suggest next action(s). For fine-grained personalized activity assistance, it is necessary to identify how an inhabitant performs the recognized type of activity in terms of its ADL profile. The discovered ADL instance can then be compared with what has already been performed to decide what need to be done next in order to accomplish the ongoing ADL.

## 5 System Implementation and Evaluation

We have implemented a feature-rich context-aware assistive system, as shown in Fig. 3. The system is developed with C# while the front-end is developed using ASP.NET with Ajax and Silverlight support for audio and graphical user experience. We use the SemWeb semantic technologies for C# [4] to create and manage semantic data in persistent storage, and use SPARQL to query persistent storage via simple graph matching. We use the Euler inference engine to implement logic-based proof mechanism for reasoning. The implemented system has been deployed in a physical kitchen environment in our SmartLab [5]. We conducted two types of experiment for evaluation purposes. The first type of experiment is aimed to evaluate the performance and accuracy of activity recognition. To do this, we design a number of activity scenarios, e.g., performing *MakeTea* activity, and then ask an actor to perform an activity following the corresponding scenario. Each time the actor uses an object, the sensor attached to the object activated. The generated sensor observations are, on one hand, collected and passed onto the system for activity recognition. On the other hand, they are manually recorded and labelled. In this way, each time a sensor is activated during the activity performance, both the system and a human evaluator can produce potential activities that might be performed by the actor. By comparing the recognition results from the system and the evaluator step by step during the performance of a designated activity scenario we are able to evaluate the accuracy of activity recognition. The second type of experiment is aimed to evaluate the applicability and robustness of the system. To do this, we used the same activity scenarios but changed the system setting using system configuration tools. Then we ask an actor to perform an identical activity twice in different system settings. We compare the recognition results from the two same-activity-scenario but different-system-setting experiments to evaluate how different system configuration can affect its performance and applicability.

All experiments have yielded desired satisfactory results demonstrating that the system is fully working and the approach is viable. The system is also evaluated by healthcare professionals from local health Trusts. From users' perspectives, they have thoroughly tested the system with very positive feedback and constructive suggestions.

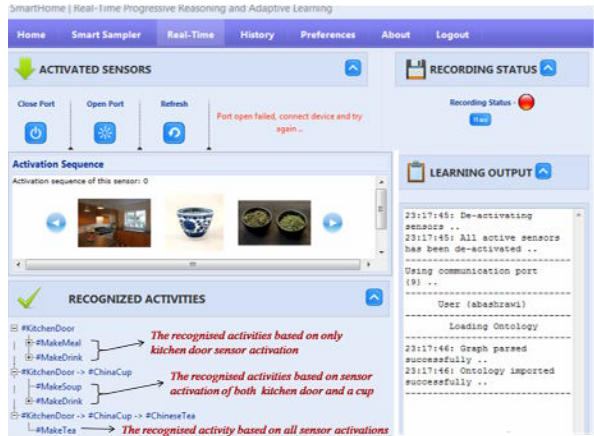


Fig. 3. The system interface in real time mode

## 6 Conclusions

In this paper we have applied ontology based knowledge engineering into the lifecycle of assistive living. We have discussed the system architecture, core functionalities, methodologies and technologies. Specifically we described the use of knowledge engineering and management for activity recognition, learning and assistance, and further detailed implementation, experiments and evaluation. Initial results have been positive and promising. While real world deployment of the system and large-scale evaluation of a diversity of use scenarios can be investigated in the future, the work has laid a solid architectural and methodological foundation.

## References

1. Chan, M., Estève, D., Escriba, C., Campo, E.: A review of smart homes—Present state and future challenges. *Computer Methods and Programs in Biomedicine* 91(1), 55–81 (2008)
2. Cook, d., Das, S.K.: How smart are our environments? An updated look at the state of the art. *Journal of Pervasive and Mobile Computing* 3(2), 53–73 (2007)
3. The Protégé, <http://protege.stanford.edu/>
4. Semantic Web/RDF Library for C#.NET, <http://razor.occams.info/code/semweb>
5. Nugent, C.D., Mulvenna, M., Hong, X.: Experiences in the Development of a Smart Lab. *The International Journal of Biomedical Engineering and Technology* 2(4), 319–331 (2009)



# Large-Scale, Exhaustive Lattice-Based Structural Auditing of SNOMED CT

Guo-Qiang Zhang

Department of Electrical Engineering and Computer Science  
Case Western Reserve University  
Cleveland, OH 44106, USA

One criterion for the well-formedness of ontologies is that their hierarchical structure form a lattice. Formal Concept Analysis (FCA) has been used as a technique for assessing the quality of ontologies, but is not scalable to large ontologies such as SNOMED CT. We developed a methodology called Lattice-based Structural Auditing (LaSA), for auditing biomedical ontologies, implemented through automated SPARQL queries, in order to exhaustively identify all non-lattice pairs in SNOMED CT. The percentage of non-lattice pairs ranges from 0 to 1.66 among the 19 SNOMED CT hierarchies. Preliminary manual inspection of a limited portion of the 518K non-lattice pairs, among over 34 million candidate pairs, revealed inconsistent use of precoordination in SNOMED CT, but also a number of false positives. Our results are consistent with those based on FCA, with the advantage that the LaSA computational pipeline is scalable and applicable to ontological systems consisting mostly of taxonomic links. This work is based on collaboration with Olivier Bodenreider from the National Library of Medicine, Bethesda, USA.

# Author Index

- Abécassis, Joël 304  
Albath, Julia 470  
Alexopoulos, Panos 388  
Anderson, Terry 378  
Anjomshoaa, Amin 412  
Awawdeh, Ruba 378
- Bai, Weijing 210  
Bashar, Abul 518  
Battiston, Roberto 366  
Bell, David 39, 101  
Bellodi, Elena 292  
Bi, Yaxin 125, 186, 506  
Black, Michaela M. 459  
Blasko, Miroslav 198
- Cai, Yi 76  
Campos, Pedro G. 270  
Capraro, Gerard T. 573  
Charnomordic, Brigitte 304  
Chaurasia, Priyanka 245  
Cheng, Peng 549  
Chen, Juan 114  
Chen, Liming 494, 609  
Coalter, Alton 482  
Cobos, Ruth 270  
Cohn, Anthony G. 1
- Dang, Yanzhong 317  
Del Vasto Terrientes, Luis 222  
Denk, Michaela 603  
Denoeux, Thierry 3  
Destercke, Sébastien 304  
Du, Jianfeng 88  
Du, Ying 494
- Fenner, Trevor 329  
Forestier, Germain 28
- Gançarski, Pierre 28  
Gärdenfors, Peter 341  
Gomes, Carlos 579  
Gröner, Gerd 51  
Grossmann, Wilfried 603  
Guo, Chonghui 234
- Hickey, Ray J. 459  
Hirasawa, Kotaro 282  
Hois, Joana 424  
Horrocks, Ian 2  
Huang, Jianping 366  
Hu, Bo 494  
Huynh, Van-Nam 160
- Jekjantuk, Nophadol 51  
Jia, Caiyan 597  
Jia, Haiyang 114  
Jiang, Yawen 597  
Jin, Zhi 555  
Johnson, Iyan 304  
Johnston, Benjamin 341
- Kaplunova, Alissa 436  
Karali, Isambo 354  
Kevitt, Paul Mc 400  
Kotis, Konstantinos 388  
Kurata, Yohei 4
- Lamma, Evelina 292  
Leopold, Jennifer L. 470, 482  
Leung, Ho-Fung 76  
Liao, Jing 186  
Li, Chungping 449  
Li, Hailin 234  
Li, Li 549  
Lin, Zongjian 561  
Liu, Dayou 114  
Liu, Jing 366  
Liu, Ming 317  
Liu, Weiru 39, 101  
Lunney, Tom 400
- Mabu, Shingo 282  
Ma, Lingling 591  
Marques, Nuno C. 63, 579  
Marrs, Gary R. 459  
McClellan, Sally 245, 518  
Miao, Chunyan 174  
Miao, Yuanqing 366  
Mirkin, Boris 329

- Moinard, Yves 585  
 Möller, Ralf 436  
 Moore, George 506  
 Moreno, Antonio 222  
 Mulholland, Paul 198  
 Muñoz, Karla 400
- Nakamori, Yoshiteru 160  
 Napoli, Amedeo 16  
 Nascimento, Susana 329  
 Nauck, Detlef 518  
 Neri, Luis 400  
 Noguez, Julieta 400  
 Nugent, Chris 186, 245, 609
- Ou, Ling 549  
 Ouyang, Xinyan 366
- Pan, Donghua 234  
 Pan, Jeff. Z. 51  
 Pappalouros, Andreas 388  
 Parr, Gerard 518  
 Patterson, David 494  
 Pereira, Luís Moniz 329  
 Perry, Kenneth 470  
 Petit, Jean-Marc 530
- Qian, Jiadong 366  
 Qi, Guilin 39, 88
- Ren, Fenghui 174  
 Riguzzi, Fabrizio 292  
 Rong, Lili 317  
 Rousset, Marie-Christine 530  
 Rózewski, Przemysław 148
- Sabharwal, Chaman L. 470  
 Sánchez, David 222  
 Scotney, Bryan 245, 518  
 Shen, Xuhui 366  
 Shen, Zhiqi 174  
 Shi, Hui 4  
 Shimada, Kaoru 282  
 Shi, Zhongzhi 543  
 Stoops, David 506  
 Szathmary, Laszlo 16
- Tahamtan, Amirreza 412  
 Tang, Lingli 591  
 Termier, Alexandre 530
- Tjoa, A. Min 412  
 The Anh, Han 63  
 Thomopoulos, Rallou 304  
 Thovex, Christophe 567  
 Tournaire, Rémi 530  
 Trichet, Francky 567
- Valtchev, Petko 16
- Wandelt, Sebastian 436  
 Wang, Hui 494, 506  
 Wang, Xinhong 591  
 Wang, Xun 256  
 Wang, Ying 101  
 Wei, Bo 555  
 Wei, Gongjin 210  
 Weippl, Edgar 412  
 Wemmert, Cédric 28  
 Wessel, Michael 436  
 Wicks, Michael C. 573  
 Wightwick, Glenn 341  
 Williams, Mary-Anne 137, 256, 341  
 Wolff, Annika 198  
 Wu, Jiangning 282  
 Wu, Shengli 125
- Xuan, Zhaoguo 317  
 Xu, Fangzhou 591
- Yang, Guangfei 282  
 Yang, Kun 543  
 Yang, Pengyi 549  
 Yan, Hongbin 160  
 Yao, Na 561  
 Yin, Meifang 210  
 Yu, Jian 597
- Zdrahal, Zdenek 198  
 Zeng, Xiaoqin 125  
 Zeren, Zhima 366  
 Zhang, Changhai 114  
 Zhang, Guo-Qiang 615  
 Zhang, Jingxiong 561  
 Zhang, Minjie 174  
 Zhang, Shuai 245  
 Zhang, Songmao 210  
 Zhang, Xuemin 366  
 Zhang, Zili 549  
 Zhao, Lili 449  
 Zowghi, Didar 555