Jeng-Shyang Pan
Shyi-Ming Chen
Ngoc Thanh Nguyen (Eds.)

LNAI 7196

Intelligent Information and Database Systems

4th Asian Conference, ACIIDS 2012
Kaohsiung, Taiwan, March 2012
Proceedings, Part I

1 Part I

CIIDS
2012

Springer

Lecture Notes in Artificial Intelligence 7196

Subseries of Lecture Notes in Computer Science

Jeng-Shyang Pan Shyi-Ming Chen
Ngoc Thanh Nguyen (Eds.)

Intelligent Information and Database Systems

4th Asian Conference, ACIIDS 2012
Kaohsiung, Taiwan, March 19-21, 2012
Proceedings, Part I

Springer

Volume Editors

Jeng-Shyang Pan
National Kaohsiung University of Applied Sciences
Department of Electronic Engineering
No. 415, Chien Kung Road, Kaohsiung 80778, Taiwan
E-mail: jengshyangpan@gmail.com

Shyi-Ming Chen
National Taichung University of Education
Graduate Institute of Educational Measurement and Statistics
No. 140, Min-Shen Road, Taichung 40306, Taiwan
E-mail: smchen@mail.ntcu.edu.tw

Ngoc Thanh Nguyen
Wrocław University of Technology, Institute of Informatics
Wybrzeże Wyspiańskiego 27, 50370, Wrocław, Poland
E-mail: ngoc-thanh.nguyen@pwr.edu.pl

ISSN 0302-9743 e-ISSN 1611-3349
ISBN 978-3-642-28486-1 e-ISBN 978-3-642-28487-8
DOI 10.1007/978-3-642-28487-8
Springer Heidelberg Dordrecht London New York

Library of Congress Control Number: 2012931775

CR Subject Classification (1998): I.2, H.3, H.2.8, H.4-5, I.4-5, K.4

LNCS Sublibrary: SL 7 – Artificial Intelligence

Typesetting: Camera-ready by author, data conversion by Scientific Publishing Services, Chennai, India

Printed on acid-free paper

Springer is part of Springer Science+Business Media (www.springer.com)

Preface

ACIIDS 2012 was the fourth event of the series of international scientific conferences for research and applications in the field of intelligent information and database systems. The aim of ACIIDS 2012 was to provide an international forum for scientific research in the technologies and applications of intelligent information, database systems and their applications. ACIIDS 2012 took place March 19–21, 2012, in Kaohsiung, Taiwan. It was co-organized by the National Kaohsiung University of Applied Sciences (Taiwan), National Taichung University of Education (Taiwan), Taiwanese Association for Consumer Electronics (TACE) and Wroclaw University of Technology (Poland), in cooperation with the University of Information Technology (Vietnam), International Society of Applied Intelligence (ISAI), and Gdynia Maritime University (Poland). ACIIDS 2009 and ACIIDS 2010 took place in Dong Hoi and Hue in Vietnam, respectively, and ACIIDS 2011 in Deagu, Korea.

We received more than 472 papers from 15 countries over the world. Each paper was peer reviewed by at least two members of the International Program Committee and International Reviewer Board. Only 161 papers with the highest quality were selected for oral presentation and publication in the three volumes of ACIIDS 2012 proceedings.

The papers included in the proceedings cover the following topics: intelligent database systems, data warehouses and data mining, natural language processing and computational linguistics, Semantic Web, social networks and recommendation systems, collaborative systems and applications, e-business and e-commerce systems, e-learning systems, information modeling and requirements engineering, information retrieval systems, intelligent agents and multi-agent systems, intelligent information systems, intelligent Internet systems, intelligent optimization techniques, object-relational DBMS, ontologies and knowledge sharing, semi-structured and XML database systems, unified modeling language and unified processes, Web services and Semantic Web, computer networks and communication systems.

Accepted and presented papers highlight new trends and challenges of intelligent information and database systems. The presenters showed how new research could lead to novel and innovative applications. We hope you will find these results useful and inspiring for your future research.

We would like to express our sincere thanks to the Honorary Chairs: Cheng-Qi Zhang (University of Technology Sydney, Australia), Szu-Wei Yang (President of National Taichung University of Education, Taiwan) and Tadeusz Wieckowski (Rector of Wroclaw University of Technology, Poland) for their support.

Our special thanks go to the Program Chairs, all Program and Reviewer Committee members and all the additional reviewers for their valuable efforts in the review process, which helped us to guarantee the highest quality of the selected papers for the conference. We cordially thank the organizers and chairs of special sessions, which essentially contribute to the success of the conference.

We would also like to express our thanks to the keynote speakers Jerzy Swiatek from Poland, Shi-Kuo Chang from the USA, Jun Wang, and Rong-Sheng Xu from China for their interesting and informative talks of world-class standard.

We cordially thank our main sponsors, National Kaohsiung University of Applied Sciences (Taiwan), National Taichung University of Education (Taiwan), Taiwanese Association for Consumer Electronics (TACE) and Wroclaw University of Technology (Poland). Our special thanks are due also to Springer for publishing the proceedings, and other sponsors for their kind support.

We wish to thank the members of the Organizing Committee for their very substantial work, especially those who played essential roles: Thou-Ho Chen, Chin-Shuih Shieh, Mong-Fong Horng and the members of the Local Organizing Committee for their excellent work.

We cordially thank all the authors for their valuable contributions and the other participants of this conference. The conference would not have been possible without their support.

Thanks are also due to the many experts who contributed to making the event a success.

Jeng-Shyang Pan
Shyi-Ming Chen
Ngoc Thanh Nguyen

Conference Organization

Honorary Chairs

Cheng-Qi Zhang	University of Technology Sydney, Australia
Szu-Wei Yang	National Taichung University of Education, Taiwan
Tadeusz Wieckowski	Wroclaw University of Technology, Poland

General Chair

Ngoc Thanh Nguyen	Wroclaw University of Technology, Poland

Program Committee Chairs

Jeng-Shyang Pan	National Kaohsiung University of Applied Sciences, Taiwan
Shyi-Ming Chen	National Taichung University of Education, Taiwan
Junzo Watada	Waseda University, Japan
Jian-Chao Zeng	Taiyuan University of Science and Technology, China

Publication Chairs

Chin-Shiuh Shieh	National Kaohsiung University of Applied Sciences, Taiwan
Li-Hsing Yen	National University of Kaohsiung, Taiwan

Invited Session Chairs

Mong-Fong Horng	National Kaohsiung University of Applied Sciences, Taiwan
Tzung-Pei Hong	National University of Kaohsiung, Taiwan
Rung-Ching Chen	Chaoyang University of Technology, Taiwan

Organizing Chair

Thou-Ho Chen	National Kaohsiung University of Applied Sciences, Taiwan

Steering Committee

Ngoc Thanh Nguyen - Chair	Wroclaw University of Technology, Poland
Bin-Yih Liao	National Kaohsiung University of Applied Sciences, Taiwan
Longbing Cao	University of Technology Sydney, Australia
Adam Grzech	Wroclaw University of Technology, Poland
Tu Bao Ho	Japan Advanced Institute of Science and Technology, Japan
Tzung-Pei Hong	National University of Kaohsiung, Taiwan
Lakhmi C. Jain	University of South Australia, Australia
Geun-Sik Jo	Inha University, Korea
Jason J. Jung	Yeungnam University, Korea
Hoai An Le-Thi	University Paul Verlaine - Metz, France
Antoni Ligęza	AGH University of Science and Technology, Poland
Toyoaki Nishida	Kyoto University, Japan
Leszek Rutkowski	Technical University of Czestochowa, Poland

Keynote Speakers

- Jerzy Swiatek

  President of Accreditation Commission of Polish Technical Universities, Poland

- Shi-Kuo Chang

Center for Parallel, Distributed and Intelligent Systems, University of Pittsburgh, USA

- Jun Wang

Computational Intelligence Laboratory in the Department of Mechanical and Automation Engineering at the Chinese University of Hong Kong, China

- Rong-Sheng Xu

Computing Center at Institute of High Energy Physics, Chinese Academy of Sciences, China

Invited Sessions Organizers

Bogdan Trawiński	Wrocław University of Technology, Poland
Oscar Cordon	Wrocław University of Technology, Poland
Przemyslaw Kazienko	Wrocław University of Technology, Poland

Ondrej Krejcar	Technical University of Ostrava, Czech Republic
Peter Brida	University of Zilina, Slovakia
Kun Chang Lee	Sungkyunkwan University, Korea
Mong-Fong Horng	National Kaohsiung University of Applied Sciences, Taiwan
Kuan-Rong Lee	Kun Shan University, Taiwan
Yau-Hwang Kuo	National Cheng Kung University, Taiwan
Wu-Chih Hu	National Penghu University of Science and Technology, Taiwan
Jeng-Shyang Pan	National Kaohsiung University of Applied Sciences, Taiwan
Shyi-Ming Chen	National Taichung University of Education, Taiwan
Chulmo Koo	Chosun University, Korea
I-Hsien Ting	National University of Kaohsiung, Taiwan
Jason J. Jung	Yeungnam University, Korea
Chaudhary Imran Sarwar	University of the Punjab, Pakistan
Tariq Mehmood Chaudhry	Professional Electrical Engineer, Pakistan
Arkadiusz Kawa	Poznan University of Economics, Poland
Paulina Golińska	Poznan University of Technology, Poland
Konrad Fuks	Poznan University of Economics, Poland
Marcin Hajdul	Institute of Logistics and Warehousing, Poland
Shih-Pang Tseng	Tajen University, Taiwan
Yuh-Chung Lin	Tajen University, Taiwan
Phan Cong Vinh	NTT University, Vietnam
Le Thi Hoai An	Paul Verlaine University, France
Pham Dinh Tao	INSA-Rouen, France
Bao Rong Chang	National University of Kaohsiung, Taiwan
Tien-Tsai Huang	Lunghwa University of Science and Technology, Taiwan

International Program Committee

Cesar Andres	Universidad Complutense de Madrid, Spain
S. Hariharan B.E.	J.J. College of Engineering and Technology Ammapettai, India
Costin Badica	University of Craiova, Romania
Youcef Baghdadi	Sultan Qaboos University, Oman
Dariusz Barbucha	Gdynia Maritime University, Poland
Stephane Bressan	NUS, Singapore
Longbing Cao	University of Technology Sydney, Australia
Frantisek Capkovic	Slovak Academy of Sciences, Slovakia
Oscar Castillo	Tijuana Institute of Technology, Mexico
Bao Rong Chang	National University of Kaohsiung, Taiwan

Hsuan-Ting Chang	National Yunlin University of Science and Technology, Taiwan
Lin-Huang Chang	National Taichung University of Education, Taiwan
Chuan-Yu Chang	National Yunlin University of Science and Technology, Taiwan
Jui-Fang Chang	National Kaohsiung University of Applied Sciences, Taiwan
Wooi Ping Cheah	Multimedia University, Malaysia
Shyi-Ming Chen	National Taichung University of Education, Taiwan
Guey-Shya Chen	National Taichung University of Education, Taiwan
Rung Ching Chen	Chaoyang University of Technology, Taiwan
Suphamit Chittayasothorn	King Mongkut's Institute of Technology, Thailand
Tzu-Fu Chiu	Aletheia University, Taiwan
Chou-Kang Chiu	National Taichung University of Education, Taiwan
Shu-Chuan Chu	Flinders University, Australia
Irek Czarnowski	Gdynia Maritime University, Poland
Ireneusz Czarnowski	Gdynia Maritime University, Poland
Jiangbo Dang	Siemens Corporate Research, USA
Tran Khanh Dang	HCMC University of Technology, Vietnam
Paul Davidsson	Malmö University, Sweden
Hui-Fang Deng	South China University of Technology, China
Phuc Do	University of Information Technology, Vietnam
Van Nhon Do	University of Information Technology, Vietnam
Manish Dixit	Madhav Institute of Technology and Science, India
Antonio F.	Murcia University, Spain
Pawel Forczmanski	West Pomeranian University of Technology, Poland
Patrick Gallinar	Université Pierre et Marie Curie, France
Mauro Gaspari	University of Bologna, Italy
Dominic Greenwood	Whitestein Technologies, Switzerland
Slimane Hammoudi	ESEO, France
Hoang Huu Hanh	Hue University, Vietnam
Jin-Kao Hao	University of Angers, France
Le Thi Hoai An	Paul Verlaine University – Metz, France
Kiem Hoang	University of Information Technology, Vietnam
Tzung-Pei Hong	National University of Kaohsiung, Taiwan
Mong-Fong Horng	National Kaohsiung University of Applied Sciences, Taiwan
Ying-Tung Hsiao	National Taipei University of Education, Taiwan

Chao-Hsing Hsu	Chienkuo Technology University, Taiwan
Wen-Lian Hsu	Academia Sinica, Taiwan
Feng-Rung Hu	National Taichung University of Education, Taiwan
Wu-Chih Hu	National Penghu University of Science and Technology, Taiwan
Hsiang-Cheh Huang	National University of Kaohsiung, Taiwan
Yung-Fa Huang	Chaoyang University of Technology, Taiwan
Tien-Tsai Huang	Lunghwa University of Science and Technology, Taiwan
Deng-Yuan Huang	Dayeh University, Taiwan
Jingshan Huang	University of South Alabama, USA
Piotr Jedrzejowicz	Gdynia Maritime University, Poland
Albert Jeng	Jinwen University of Science and Technology, Taiwan
Alcala-Fdez Jesus	University of Granada, Spain
Jason J. Jung	Yeungnam University, South Korea
Janusz Kacprzyk	Polish Academy of Sciences, Poland
Radosaw Piotr Katarzyniak	Wroclaw University of Technology, Poland
Muhammad Khurram Khan	King Saud University, Saudi Arabia
Cheonshik Kim	Sejong University, Korea
Joanna Kolodziej	University of Bielsko-Biala, Poland
Ondrej Krejcar	VSB - Technical University of Ostrava, Czech Republic
Dariusz Krol	Wroclaw University of Technology, Poland
Wei-Chi Ku	National Taichung University of Education, Taiwan
Tomasz Kubik	Wroclaw University of Technology, Poland
Bor-Chen Kuo	National Taichung University of Education, Taiwan
Kazuhiro Kuwabara	Ritsumeikan University, Japan
Raymond Y.K. Lau	City University of Hong Kong, Hong Kong
Kun Chang Lee	Sungkyunkwan University, Korea
Chin-Feng Lee	Chaoyang University of Technology, Taiwan
Eun-Ser Lee	Andong National University, Korea
Huey-Ming Lee	Chinese Culture University, Taiwan
Chunshien Li	National Central University, Taiwan
Tsai-Hsiu Lin	National Taichung University of Education, Taiwan
Yuan-Horng Lin	National Taichung University of Education, Taiwan
Chia-Chen Lin	Providence University, Taiwan
Hao-Wen Lin	Harbin Institute of Technology, China
Min-Ray Lin	National Taichung University of Education, Taiwan
Hsiang-Chuan Liu	Asia University, Taiwan

Yu-lung Lo	Chaoyang University of Technology, Taiwan
Ching-Sung Lu	Tajen University, Taiwan
James J. Lu	Emory University, USA
Janusz Marecki	IBM T.J. Watson Research Center, USA
Vuong Ngo Minh	Ho Chi Minh City University of Technology, Vietnam
Tadeusz Morzy	Poznan University of Technology, Poland
Kazumi Nakamatsu	School of Human Science and Environment, University of Hyogo, Japan
Grzegorz J. Nalepa	AGH University of Science and Technology, Poland
Jean-Christophe Nebel	Kingston University, UK
Vinh Nguyen	Monash University, Australia
Manuel Nunez	Universidad Complutense de Madrid, Spain
Marcin Paprzycki	Systems Research Institute of the Polish Academy of Sciences, Poland
Witold Pedrycz	University of Alberta, Canada
Ibrahima Sakho	University of Metz, France
Victor Rung-Lin Shen	National Taipei University, Taiwan
Tian-Wei Sheu	National Taichung University of Education, Taiwan
Chin-Shiuh Shieh	National Kaohsiung University of Applied Sciences, Taiwan
Shu-Chuan Shih	National Taichung University of Education, Taiwan
An-Zen Shih	Jinwen University of Science and Technology, Taiwan
Gomez Skarmeta	Murcia University, Spain
Serge Stinckwich	IRD, France
Pham Dinh Tao	National Institute for Applied Sciences Roue, France
Wojciech Thomas	Wroclaw University of Technology, Poland
Geetam Singh Tomar	Gwalior Malwa Institute of Technology and Management, India
Dinh Khang Tran	School of Information and Communication Technology HUST, Vietnam
Bogdan Trawiski	Wrocaw University of Technology, Poland
Hoang Hon Trinh	Ho Chi Minh City University of Technology, Vietnam
Hong-Linh Truong	Vienna University of Technology, Austria
Chun-Wei Tseng	Cheng Shiu University, Taiwan
Kuo-Kun Tseng	Harbin Institute of Technology, China
Felea Victor	Alexandru Ioan Cuza University of Iai, Romania
Phan Cong	Vinh, NTT University, Vietnam
Yongli Wang	North China Electric Power University, China

Lee-Min Wei	National Taichung University of Education, Taiwan
Michal Wozniak	Wroclaw University of Technology, Poland
Homer C. Wu	National Taichung University of Education, Taiwan
Xin-She Yang	National Physical Laboratory, UK
Horng-Chang Yang	National Taitung University, Taiwan
Shu-Chin Yen	Wenzao Ursuline College of Languages, Taiwan
Ho Ye	Xidian University, China

Table of Contents – Part I

Agent System

Intelligent Systems(1)

Intelligent Systems(2)

Intelligent Systems(3)

Multiple Model Approach to Machine Learning(1)

Multiple Model Approach to Machine Learning(2)

Intelligent Supply Chains

Table of Contents – Part II

Clustering Technology

Intelligent Digital Watermarking and Image Processing

Intelligent Management of e-Business

Intelligent Media Processing

Modelling and Optimization Techniques in Information Systems, Database Systems and Industrial Systems

User Adaptive Systems(1)

User Adaptive Systems(2)

Advances in Nature-Inspired AutonQomic Computing and Networking

Table of Contents – Part III

Ubiquitous Decision Support with Bayesian Network and Computational Creativity

Computational Intelligence

Human Computer Interaction

Innovation in Cloud Computing Technology and Application

Innovative Computing Technology

Intelligent Service

Intelligent Signal Processing and Application

A Multi-agent Strategy for Integration of Imprecise Descriptions

Grzegorz Skorupa, Wojciech Lorkiewicz, and Radosław Katarzyniak

Wroclaw University of Technology
{grzegorz.skorupa,wojciech.lorkiewicz,radoslaw.katarzyniak}@pwr.wroc.pl

Abstract. One of the fundamental challenges of distributed multi-agent systems relates to the problem of knowledge integration – where a collective stance of the distributed system needs to be determined and justified. Assuming a simple multi-agent system we present an intuitive and consistent approach to the integration of imprecise descriptions of individual beliefs about the external world. In particular, focusing our attention to simple modal statements of certainty, i.e., possibility, belief and knowledge, we state rational set of common-sense postulates against the integration results. Further, utilising the Grounding Theory, as the means for appropriate grounding of statements, and incorporating the approach with theory of mind, as the underlying mechanism for interpretation of statements, we introduce a two-stage integration procedure. Finally, we prove the appropriateness of the proposed solution, show its basic properties, and provide a simple computational example.

Keywords: knowledge integration, consensus, multi-agent system, modal statements.

1 Introduction

Distributed systems are recently gaining interest and their widespread usage is becoming more and more popular. They usually provide an effective and a low cost solution to problems that are often unmanageable for a monolithic and centralised approach. Unfortunately, allowing separate autonomous components to coordinate, exchange knowledge and maintain relations is an exquisitely demanding task, especially in dynamic environments. One of such fundamental tasks relates to knowledge integration, where a collective stance of the distributed system needs to be determined and justified. In a setting where the agents are additionally highly autonomous, i.e. can represent inconsistent or even opposing views, resolving the collective knowledge state is a notoriously hard problem [9].

Here we follow the cognitive linguistics approach to communication[1] and adopt the phenomenological stance, where all individuals maintain a conscious and intentional relationship to the external world through their bodily experiences. Incorporating the Grounding Theory model [6–8], the autonomous agent through

[1] As opposed to classical definitions of categories, the existence of a mind-independent reality (objectivist realism) and absolute truths.

J.-S. Pan, S.-M. Chen, N.T. Nguyen (Eds.): ACIIDS 2012, Part I, LNAI 7196, pp. 1–10, 2012.

the interaction with the external environment experiences surface mode qualia (understood as "raw feels"), that trigger a particular cognitive schema within the individual. In short, an observation is an individual perception that is introduced into the agent body and represented as embodied structures. Namely, it is assumed that the cognitive agent can store internally reflections of perceptually available states of properties P in perceptually available objects O. It means that each aspect of the external world is recognisable for the agent and can become a part of the body of its empirically originated knowledge.

The internal organisation of the agent is strictly private and individual, consequently the embodied structures are not shared among the interacting agents and cannot be directly communicated. As such, for an agent to share its current viewpoint, it is necessary to utilise a language that is established within the population. In particular, registering a particular language symbol triggers a consistent, with the knowledge stance of the speaker, reaction within the hearer. Additionally, following the view of Lakoff and Johnson[2], agent's internal system of accessible concepts is strictly related to the structure of the external world, i.e. the language symbol is grounded in the embodied experience [4].

According to Dennet multiple 'exposure to x – that is, sensory confrontation with x over suitable period of time – is the normally sufficient condition for knowing (or having true beliefs) about x'[1]. Based on this idea, the Grounding Theory [6–8] defines a mechanism for an agent to fill in the unobserved parts of the environment with cognitive schema extracted from the past empirical experiences (corresponding with a particular state of an unobserved property in a given object). Moreover, observation in which an agent observed an object exhibiting a property (not exhibiting a property) makes a corresponding cognitive scheme stronger (weaker) in relation to the complementary schemes. These strengths, called relative grounding strengths (λ), along with a system of modality thresholds (for modal levels of possibility, belief and knowledge), serve as means for grounding appropriate modal statements.

Obviously, the agents do not share the same experiences, i.e., their occasional observations of the environment do not have to be synchronised nor have to focus on the same aspects of the external world. As such grounded modal statements, which are further uttered by the agents, do not have to be identical, even more – do not have to be consistent. Consequently, determining the collective stance of the population is not a trivial task and a system capable of determining such a global view requires a specialised integration mechanism.

It should be noted that there exists a vast literature on the topic of knowledge integration, in particular in belief integration tasks [2, 3]. The most prominent and well known is the research concerning belief revision, i.e. researching the problem of assessing new pieces of information and how to incorporate them in already established knowledge bases, belief merging, i.e. researching the problem of integrating already established knowledge bases, and voting, i.e. researching the means for a collective compromise between multiple preferences (knowledge pieces). However,

[2] '(...) Human language and thought are structured by, and bound to, an embodied experience (...)' [5].

applying the mechanisms from the aforementioned studies leads to several problems and questions. In particular, as a consequence that they neglect the fact that modal statements should be grounded in an individual agent. *Each such grounded modal statement represents a particular cognitive stance of the uttering agent and as such the integration task should take this cognitive stance into account.* Encouraged by this gap in the literature, in the following sections, we formulate an alternative approach to the integration task – incorporating the mechanisms of Grounding Theory.

2 Knowledge Integration Task

We assume a multi-agent system comprised of population of multiple observing agents $\mathcal{A} = \{1, 2, \ldots, A\}$. Each such agent $a \in \mathcal{A}$ is occasionally interacting with the environment, when it stores binary (exhibiting or not exhibiting) states of distinguished properties $p \in P$ in currently available objects $o \in O$. Further, based on past experiences, i.e. utilising the Grounding Theory mechanism, an agent is able to utter it's individual view (cognitive stance) about the state of external environment through a set of modal statements of possibility, believe and knowledge. As aforementioned, each grounded formula ω represents agent's a individual certainty concerning the true nature of the state of the external environment. Moreover, the set of *all* possible formulae $\Omega^{(a)}(t)$, i.e. modal statements that can be grounded in a particular situation $t \in T$, reflects exhaustively agent's a current stance.

2.1 Task Definition

In the proposed approach the integration is realised through a distinguished integrating agent $\hat{a}$, which sole task is to provide an integrated view of the current stance of the collective (collective stance). This fundamental functionality is realised by the means of a two stage integration process – *reflection* and *integration.*

First, the integrating agent $\hat{a}$ collects all modal statements grounded by observing agents $\mathcal{A}$. In particular, it queries the population for descriptions concerning a particular pair of properties $p, q \in P$ in a given object $o \in O$.[3] Following the Grounding Theory, i.e., the model of grounding simple modalities and modal conjunctions, we limit the basic formulae to modal statements of agent's certainty about a particular state $\phi \in \{p, q, \neg p, \neg q, p \wedge q, p \wedge \neg q, \neg p \wedge q, \neg p \wedge \neg q\}$ of the external world, i.e. assuming three modal levels of possibility ($Pos(\phi)$), belief ($Bel(\phi)$), and knowledge ($Know(\phi)$). As such, each queried agent a responds by sending its individual beliefs $\Omega^{(a)}$ (See def. 1), i.e. an exhaustive set of *all* grounded modal statements.

[3] For the sake of simplicity, we further omit the object from the notation, i.e. we use p instead of $p(o)$, and assume that at a certain point of time $t_I \in T$ the integration task is strictly limited to a given pair of properties $q, p \in P$.

Definition 1. *Let $\Omega^{(a)}$ denote a complete set of grounded statements concerning properties $p, q \in P$ and received from agent a. More generally, let:*

$$\Omega = \{\Omega^{(1)}, \Omega^{(2)}, ..., \Omega^{(A)}\} \tag{1}$$

denote all sets of statements received from agents $1, 2, ..., A$ respectively.

Remark 1. Grounding Theory provides the means for generating a consistent set of grounded statements. In particular, it defines the proper binding between the modal level of certainty and the internal cognitive stance of the agent. This binding is implemented through a system of modality thresholds over simple statistics, and guarantees that statements fulfil the common-sense requirements and are mutually consistent. For example, if an agent is able to state that it is certain that $Know(p \wedge q)$, then it is impossible to state $Pos(\phi)$ of the complimentary states ($\phi \in \{p \wedge \neg q, \neg p \wedge q, \neg p \wedge \neg q\}$), but when it only believes that $Bel(p \wedge q)$ then it may also state $Pos(\phi)$.[4]

The cognitive stance is agent's mental representation of a world and can be represented as a vector $\lambda = (\lambda_{p\wedge q}, \lambda_{p\wedge\neg q}, \lambda_{\neg p\wedge q}, \lambda_{\neg p\wedge\neg q})$. Each coordinate represents a subjective strength of a particular cognitive scheme (grounding set) where object o possessed or lacked particular features. For example value of $\lambda_{p\wedge q}$ corresponds to how probable is a situation where object o possessed both, p and q, features. The higher the value of $\lambda_{p\wedge q}$ the greater the chance of o being both p and q. It is assumed that cognitive state meets following general constraints:

$$\lambda_{p\wedge q} + \lambda_{p\wedge\neg q} + \lambda_{\neg p\wedge q} + \lambda_{\neg p\wedge\neg q} = 1 \tag{2}$$

$$0 \leq \lambda_{p\wedge q} \leq 1,\ 0 \leq \lambda_{p\wedge\neg q} \leq 1,\ 0 \leq \lambda_{\neg p\wedge q} \leq 1,\ 0 \leq \lambda_{\neg p\wedge\neg q} \leq 1 \tag{3}$$

Remark 2. It is assumed that each observing agent a provides its complete knowledge concerning the given pair of properties. In particular, not uttering a statement (concerning a particular combination of properties, e.g. $p \wedge q$) is equivalent to a situation where agent's beliefs (concerning that combination) are not strong enough to ground a statement of possibility (e.g. $Pos(p \wedge q)$), believe (e.g. $Bel(p \wedge q)$) or knowledge (e.g. $Know(p \wedge q)$).

Further, still in the reflection stage, received sets of statements are analysed individually. Applying the intentional approach, i.e. assuming that the uttered modal statements $\Omega^{(a)}$ represent a particular cognitive viewpoint of the observer a, the integrating agent creates it's internal reflection. Analogous to the theory of mind, the set of modal statements is projected onto the internal structures of the agent that mirror, or engage in, a consistent cognitive stance that would trigger the integrating agent to ground all of the modal statements from $\Omega^{(a)}$. In essence, internal reflection is realised by determining the relative grounding strengths that result in a particular cognitive stance consistent with the

[4] For precise definition of the notions of grounding set, modality thresholds, grounding strength and their internal relations, please see [6–8].

observer's viewpoint. Consequently, the integrating agent speratelly translates all of the received distributed stances $\Omega^{(i)}$ to its individual internal reflections $R(\Omega^{(i)}) = \lambda^{(i)}$, namely a reflection of its cognitive state:

Definition 2. *For a given observer agent a and it's complete set of grounded statements $\Omega^{(a)}$, let $R : \Omega \to [0,1]^4$ denote the internal reflection function that determines the internal reflection $R(\Omega^{(a)}) = \lambda^{(a)} = (\lambda_{p\wedge q}, \lambda_{p\wedge\neg q}, \lambda_{\neg p\wedge q}, \lambda_{\neg p\wedge\neg q})$ of the a's cognitive stance. In general, let set $\Lambda = \{\lambda^{(1)}, \lambda^{(2)}, ..., \lambda^{(A)}\}$ denote the set of all reflections of cognitive states.*

Next we enter the integration stage, where the internal reflections Λ of gathered stances Ω are further integrated to form a single consistent knowledge stance (See sec. 3). In particular, based on the integrated reflection the agent determines (See sec. 2.1) a collective reflection, i.e. creates an integrated cognitive state $\lambda^* = (\lambda^*_{p\wedge q}, \lambda^*_{p\wedge\neg q}, \lambda^*_{\neg p\wedge q}, \lambda^*_{\neg p\wedge\neg q})$ of community of agents. Method of constructing vector λ^* ensures that required properties (See sec.2.2) are met. Finally, the integrated cognitive state λ^* that serves as the source for the integrated set of modal statements Ω^* – describing agents' community knowledge.

2.2 Integration Task Common-Sense Requirements

In general, the integrating agent is searching for a consistent set of statements Ω^* that in the best way describes the collective stance of the entire population. Best meaning that the integration process can be characterised by a common-sense set of postulates (inspired by [9]):

Identity Integration. The integration of statements originated from a single agent should result in an unchanged set of statements.

Superiority of Knowledge. If one agent states its strict certainty, i.e. $Know(\phi)$, and none of the agents states a contradictory statement with the same level of certainty, then the integrated set of statements should include this strict certainty.

Superiority of Community. If there are no strict certainty statements (with $Know$ operator), and additionally if N agents stated their belief concerning ϕ, i.e. $Bel(\phi)$, and M agents stated the possibility of ϕ, i.e. $Pos(\phi)$, then the integrated set of statements should maintain $Bel(\phi)$ if $N \gg M$ or $Pos(\phi)$ if $N \ll M$.

Statement Presence. If there are no strict certainty statements (with $Know$ operator), statements that are present in all input sets $\Omega^{(a)}$ should be included in the integrated set of statements.

Inconsistency of Knowledge. If one of the agents states its certainty concerning ϕ, i.e. $Know(\phi)$, and there exists an agent that states its certainty concerning $\neg\phi$, then none of this modal statements should be included in the integrated set of statements.

3 Integration Strategy

We present a formal description of the integration process – reflection stage (method of collecting and determining the individual reflections of knowledge states $\lambda^{(a)}$) and integration stage (method of determining the collective knowledge state λ^* and translating to integrated set of modal statements Ω^*).

3.1 Reflection of Agent's Cognitive State

Grounding Theory [6–8] defines a set of constraints imposed on agent's cognitive state λ. These constraints define which statements can and can't be used to describe observed situation. Constraints, i.e. a system of modality thresholds, for modal conjunctions and simple statements are presented in table 1.[5]

Table 1. Constraints for modal statements

statement on $p \wedge q$	constraint on cognitive state
$Know(p \wedge q)$	$\lambda_{p\wedge q} = 1$
$Bel(p \wedge q)$	$\lambda_B^k \leq \lambda_{p\wedge q} < 1$
$Pos(p \wedge q)$	$\lambda_P^k < \lambda_{p\wedge q} < \lambda_B^k$
no message about $p \wedge q$	$\lambda_{p\wedge q} \leq \lambda_P^k$
statement on p	constraint on cognitive state
$Know(p)$	$\lambda_{p\wedge q} + \lambda_{p\wedge\neg q} = 1$
$Bel(p)$	$\lambda_B^s \leq \lambda_{p\wedge q} + \lambda_{p\wedge\neg q} < 1$
$Pos(p)$	$\lambda_P^s < \lambda_{p\wedge q} + \lambda_{p\wedge\neg q} < \lambda_B^s$
no message about p	$\lambda_{p\wedge q} + \lambda_{p\wedge\neg q} \leq \lambda_P^s$

For example a cognitive state $\lambda = (1, 0, 0, 0)$ allows statements $Know(p \wedge q), Know(p), Know(q)$ but does not allow any statement on $\neg p \wedge \neg q$. A cognitive state $\lambda = (0.8, 0.2, 0, 0)$ allows $Bel(p \wedge q), Bel(p), Pos(p \wedge \neg q)$ and a few others.

Each agent communicates all grounded statements (complete set $\Omega^{(a)}$) and all agents are rational (constraints are not contradictory), as such each statement (or lack of statement) from $\Omega^{(a)}$ denotes particular constraints (See tab.1). Let Ξ denote a set of constraints resulting from agent's a statements $\Omega^{(a)}$ and general constraints (See eq.2-3). Consequently, points that satisfy all of the constraints represent a cognitive state where only the statements from $\Omega^{(a)}$ can be grounded.

The task of the first stage of the integration process, i.e. reflection stage, is to find the internal reflection $\lambda = R(\Omega^{(a)})$ of agent's a cognitive state. In general, the intermediate task is to find such a point $\lambda = (\lambda_{p\wedge q}, \lambda_{p\wedge\neg q}, \lambda_{\neg p\wedge q}, \lambda_{\neg p\wedge\neg q})$ for every agent.

In geometrical terms, constraints form at most three-dimensional figure in four-dimensional space. It can be a single point (when $Know(p \wedge q)$ is said), a

[5] Constraints for modal messages on q, $\neg p$, $\neg q$, $p \wedge \neg q$, $\neg p \wedge q$ and $\neg p \wedge \neg q$ are defined analogously. Parameters λ_B^k, λ_P^k, λ_P^s and λ_B^s are set so that they meet requirements from [6–8].

line segment ($Know(p)$ is said) or a tetrahedron (no '$Know$' statements are said) – every interior point meets the constraints Ξ. However, due to the restrictions imposed by strict inequalities this property is not met by all of the figure's boundary points . It should be noted, that when the figure is degenerated to a single point it must satisfy all of the constraints Ξ. Additionally, each non-boundary point of a line-segment must satisfy all of the constraints Ξ.

In order to find a point satisfying constraints Ξ we find all of the figure's (defined by Ξ) vertices. Let:

$$V = \{v^{(1)}, v^{(2)}, ..., v^{(C)}\}, \text{ where } v^{(c)} = (v_1^{(c)}, v_2^{(c)}, v_3^{(c)}, v_4^{(c)}) \tag{4}$$

denote a set of vertices of figure defined by constraints Ξ.[6]

Found vertices are used to calculate a single point lying within the figure's interior. Point λ is called a reflection of agent's cognitive state and is calculated as an average of vertices:

$$\lambda = (\lambda_{p\wedge q}, \lambda_{p\wedge\neg q}, \lambda_{\neg p\wedge q}, \lambda_{\neg p\wedge\neg q}) = \frac{1}{C}\sum_{c=1}^{C} v^{(c)} \tag{5}$$

The figure defined by Ξ is convex, in each case, but not strongly convex. This ensures that the point λ (See eq.5) lies within the figure's bounds. If the figure is reduced to a single point, then λ is exactly that point. If the figure is reduced to a line segment, then λ lies at the middle of that segment. In case the figure has a three-dimensional shape, then λ lies within its interior. Consequently, the point λ meets all of the criteria given by Ξ.

Vertices V can be obtained using well-known algebraic properties of 4 dimensional Euclidean space and simple linear equations. For the sake of simplicity we modify the set Ξ to define two auxiliary sets $\Xi_{\leq}, \Xi_{=}$ of changed constraints. Consequently we define a set $\Xi_{\leq}$, which contains only linear equations and soft inequalities, and a set $\Xi_{=}$, which contains only linear equations.

To obtain set $\Xi_{\leq}$ we change every equality constraint from set Ξ so that it has a form: $a_1\lambda_{p\wedge q} + a_2\lambda_{p\wedge\neg q} + a_3\lambda_{\neg p\wedge q} + a_4\lambda_{\neg p\wedge\neg q} + a_5 = 0$ and change every inequality constraint so that it has form $a_1\lambda_{p\wedge q}+a_2\lambda_{p\wedge\neg q}+a_3\lambda_{\neg p\wedge q}+a_4\lambda_{\neg p\wedge\neg q}+a_5 \leq 0$. For example, a constraint for $Bel(p \wedge q)$ can be transformed into two constraints of the form $B_k - \lambda_{p\wedge q} \leq 0$ and $\lambda_{p\wedge q} - 1 \leq 0$.

To create set $\Xi_{=}$ we change every inequality constraint from $\Xi_{\leq}$ to equality. For example a constraint for statement $Pos(p)$ can be transformed into $\lambda_{p\wedge q} + \lambda_{p\wedge\neg q} - \lambda_P^s = 0$.

Let $A = [a_{ij}]$ be a $5 \times card(\Xi_{=})$ matrix. Holding parameters a_{ij} from $\Xi_{=}$ constraints. Conditions for constraints from set $\Xi_{=}$ can be written as a system of linear equations of the matrix form:

$$[\lambda\ 1]A = [\lambda_{p\wedge q}\ \lambda_{p\wedge\neg q}\ \lambda_{\neg p\wedge q}\ \lambda_{\neg p\wedge\neg q}\ 1]A = 0 \tag{6}$$

Equation 6 always has no solutions – however, every minor of 5×5 submatrix of A is a good candidate for a figure vertex.

[6] The number (C) of vertices depends on the figure's shape.

Theorem 1. *Point v is a vertex of figure defined by constraints $\Xi_{\leq}$ iff: 1) There exists submatrix A' of size 5×5 created from matrix A such that $det(A') \neq 0$ and $[v\ 1]A' = 0$. 2) Vertex v satisfies all constraints from set $\Xi_{\leq}$.*

Following the theorem 1 one can find all vertices of figure defined by constraints $\Xi_{\leq}$ by solving equations $[v\ 1]A' = 0$ for every submatrix A' of A and checking result against $\Xi_{\leq}$. This way all vertices V defined in 4 are found. Later vertices V are used to calculate agent cognitive state λ using equation 5.

3.2 Integration Procedure

Sentences $\Omega^{(a)}$ of agent a are transformed to a reflection of cognitive state using method described in section 3.1. Doing so for every agent results in a set of reflections of cognitive states $\Lambda = (\lambda^{(1)}, \lambda^{(2)}, ..., \lambda^{(A)})$. Cognitive states need to be integrated into one state λ^* describing the entire community of agents.

The integration of Λ to λ^* must meet properties described in section 2.2. In particular it has to meet *superiority of knowledge* and *inconsistency of knowledge* requirements. In order to do so, statements with '*Know*' operator must be handled separately. Let us denote three types T1, T2, T3 of vectors $\lambda^{(a)}$, and the resultant important properties P1, P2, P3:

T1 Vector $\lambda^{(a)}$ is T1 if it has one coordinate equal to 1 and all other coordinates equal to 0 (for example $(1, 0, 0, 0)$).

T2 Vector $\lambda^{(a)}$ is T2 if it has two non-zero coordinates and two coordinates equal to 0 (for example $(0.8, 0.2, 0, 0)$).

T3 Vector $\lambda^{(a)}$ is T3 if it is neither T1, nor T2.

P1 If '*Know*' conjunction was said by a, vector is of type T1.

P2 If not P1 and *Know* simple statement was said, λ is of type T2.

P3 Vectors $\lambda^{(1)}, \lambda^{(2)}$ are contradictory if one has zeros at every position the other one has non-zero values; for example (1,0,0,0) and (0,0.7,0,0.3).

To handle knowledge statements, within Λ we search for every vector T1 or T2 and put it into set Λ_{temp}. Next, using P3, we eliminate contradictory vectors from Λ_{temp} to form a set Λ_k. Resulting set Λ_k contains non-contradictory vectors for whom statements with '*Know*' operator were said. There are 4 possible situations: S1) Λ_k contains a vector of type T1. S2) Λ_k contains only vectors T2 where exactly *the same* coordinates are non-zero. S3) Λ_k contains many vectors T2 with two but not necessary the same non-zero coordinates. S4) Λ_k is empty.

Cases S1 and S3. Integration result λ^* is of type T1 where coordinate with value 1 is determined based on Λ_k.

Case S2. Integration result λ^* is of type T2 where coordinates with non-zero values are determined based on Λ_k. Assume vector is of the form $(\alpha, 1-\alpha, 0, 0)$, where $\alpha \in (0, 1)$ is searched parameter. Then solution λ^* is calculated as:

$$\lambda^* \leftarrow \min_{\lambda=(\alpha,1-\alpha,0,0)} \sum_{a=1}^{A} dist(\lambda, \lambda^{(a)})^2, \quad \alpha \in (0, 1) \tag{7}$$

Case S4. Solution λ^* is of type T3 calculated using:

$$\lambda^* \;\leftarrow\; \min_{\lambda} \sum_{a=1}^{A} dist(\lambda, \lambda^{(a)})^2 \tag{8}$$

Function *dist* is euclidean distance. Equations 7 and 8 can be analytically solved. Equation 8 leads to a solution that is simply an average of vectors from $\mathit{\Lambda}$.

In the end, the resulting vector λ^* is translated back to modal statements $\mathit{\Omega}^*$ using constraints from table 1.

3.3 Integration Properties

Our integration strategy meets all of the stated requirements (See sec. 2.2). *Identity integration* is guaranteed due to the method of calculating $\lambda^{(a)}$. *Superiority of knowledge* is met, because of special handling of cases S1, S2 and S3. *Inconsistency of knowledge* is met because of the method of constructing set $\mathit{\Lambda}_k$. *Superiority of community* is met due to solution of equation 8. This solution is an average over coordinates. When N is much greater than M resulting vector is strongly pushed towards '*Bel*'. In case $N \ll M$ otherwise. Because solution is an average over coordinates also *statement presence* is met. Result as an average value must meet any inequality constraints met by all input vectors $\mathit{\Lambda}$.

Table 2. Exemplary input sentences and integrated output sentences

Example 1	
agent 1	$Know(p),\ Bel(q),\ Bel(p \wedge q),\ Pos(\neg q),\ Pos(p \wedge \neg q)$
agent 2	$Know(q),\ Bel(p),\ Bel(p \wedge q),\ Pos(\neg p),\ Pos(\neg p \wedge q)$
agent 3	$Bel(p),\ Pos(\neg p),\ Pos(q),\ Pos(\neg q),\ Pos(p \wedge q),\ Pos(p \wedge \neg q)$
result	$Know(p),\ Know(q),\ Know(p \wedge q)$
Example 2	
agent 1	$Know(\neg p),\ Know(q),\ Know(\neg p \wedge q)$
agent 2	$Know(p),\ Bel(q),\ Bel(p \wedge q),\ Pos(\neg q),\ Pos(p \wedge \neg q)$
agent 3	$Bel(q),\ Pos(p),\ Pos(\neg p), Pos(\neg q),\ Pos(p \wedge q),\ Pos(\neg p \wedge q)$
result	$Know(q),\ Pos(p),\ Pos(\neg p),\ Pos(p \wedge q),\ Pos(\neg p \wedge q)$
Example 3	
agent 1	$Bel(p),\ Bel(q),\ Bel(p \wedge q),\ Pos(\neg q),\ Pos(p \wedge \neg q)$
agent 2	$Bel(p),\ Pos(\neg p),\ Pos(q),\ Pos(\neg q),\ Pos(p \wedge q), Pos(p \wedge \neg q), Pos(\neg p \wedge \neg q)$
agent 3	$Bel(p),\ Pos(q),\ Pos(\neg q),\ Pos(p \wedge q),\ Pos(p \wedge \neg q)$
result	$Bel(p),\ Pos(\neg p),\ Pos(q),\ Pos(\neg q),\ Pos(p \wedge q),\ Pos(p \wedge \neg q)$

4 Computational Example

Table 2 contains three examples of the integration process, all in a setting of 3 observing agents. Example 1 shows how *superiority of knowledge* is sustained.

One agent knows that p and second knows that q. After integration we know both p and q. Within example 2 there are contradictory statements $Know(\neg p \wedge q)$, $Know(\neg p)$ in agent 1 and $Know(p)$ in agent 2. Integrating agent resigns from knowledge of p but stays with knowledge of q. Example 3 contains no knowledge statements. Result is a balanced response. Sentences $Bel(p)$, $Pos(\neg q)$, $Pos(p \wedge \neg q)$ uttered by every agent are also present in result.

5 Conclusions

Presented research provides an intuitive and consistent approach to the problem of integrating distributed and imprecise descriptions of the external world. Approach that seems ideal to allow a system of comprised of distributed autonomous agents to operate collectively in dynamic environments. At first, we managed to provide a general architecture of a multi-agent system capable of generating collective descriptions of internally distributed knowledge stances. Further, focusing our attention to modal descriptions, that express agent's individual certainty about the state of the external world, we formulate fundamental set of common-sense postulates against the resultant integrated stance.

Utilising the Grounding Theory, as the means for appropriate grounding of modal statements, and incorporating the theory of mind approach, as the underlying mechanism for interpreting modal statements, we proposed and analysed a two-stage integration procedure. Finally, we proved the appropriateness of the proposed solution and provided a few computational examples.

Acknowledgements. This paper was partially supported by Grant no. N N519 444939 funded by Polish Ministry of Science and Higher Education (2010-2013).

References

1. Dennett, D.C.: True believers: The intentional strategy and why it works. In: Stich, S.P., Warfield, T.A. (eds.) Mental Representation: A Reader. Blackwell (1994)
2. Gardenfors, P.: Belief Revision. Cambridge University Press, New York (1999)
3. Hansson, S.O.: A Survey of Non-Prioritized Belief Revision. Erkenntnis 50(2/3), 413–427 (1999)
4. Harnad, S.: The Symbol Grounding Problem. Physica D 42, 335–346 (1990)
5. Lakoff, G., Johnson, M.: Philosophy In The Flesh: the Embodied Mind and its Challenge to Western Thought. Basic Books (1999)
6. Katarzyniak, R.: Grounding Atom Formulas and Simple Modalities in Communicative Agents. Applied Informatics, 388–392 (2003)
7. Katarzyniak, R.: The Language Grounding Problem and its Relation to the Internal Structure of Cognitive Agents. J. UCS 11(2), 357–374 (2005)
8. Katarzyniak, R.: On some properties of grounding uniform sets of modal conjunctions. Journal of Intelligent and Fuzzy Systems 17(3), 209–218 (2006)
9. Nguyen, N.T.: Consensus systems for conflict solving in distributed systems. Information Sciences 147(1-4), 91–122 (2002)

Performance Evaluation of Multiagent-System Oriented Models for Efficient Power System Topology Verification

Kazimierz Wilkosz, Zofia Kruczkiewicz, and Tomasz Babczyński

Wrocław University of Technology Wybrzeże Wyspiańskiego 27,
50-370 Wrocław, Poland
{Kazimierz.Wilkosz,Zofia.Kruczkiewicz,
Tomasz.Babczynski}@pwr.wroc.pl

Abstract. In the paper, the power system topology verification with use of multiagent systems is considered. Two multiagent systems are taken into account. One of these systems is a modification of the second one. In the modified system, there are additionally so-called substation agents. Stages of analysis, design and investigation of performance characteristics of presented multiagent systems are described in the paper. The goal of the paper is presentation of performance characteristics of the mentioned multiagent systems in terms of probabilistic characteristics of agent activity and created messages. The carried out investigations show that the modified multiagent system has much better features than the earlier system.

Keywords: multiagent system, interaction protocol, power system topology verification, performance evaluation.

1 Introduction

Knowledge of the correct connectivity model of a Power System (PS), i.e. a correct PS topology model is very essential from the point of view of a monitoring of PS. In this context, not only possession of proper procedure of building a PS topology model but also possession of effective procedure for verification of this model is very important, especially when building a PS topology model and its verification should be realized automatically. In this paper, the verification of the PS topology model is considered. It is assumed that Topology Verification (TV) is performed with the use of the method described in [1]. The original method from [1] decomposes TV for the whole PS into many Local TVs (LTVs). The paper [2] presents idea of utilization of Agent Technology (AT) for PS TV realized with the use of the method from [1]. Three different solutions of Multi-Agent Systems (MASs) for PS TV are considered in [3]. It was found that such MAS, which initializes PS TV when introductory analysis of measurement data can point out occurrence of any error in a PS topology model (i.e. Topology Error – TE), is the

J.-S. Pan, S.-M. Chen, N.T. Nguyen (Eds.): ACIIDS 2012, Part I, LNAI 7196, pp. 11–20, 2012.

most advantageous. In [4], the new solution of MAS for PS TV is introduced. In [4], not only existence of one electrical node in one substation but also existence of several electrical nodes in one substation is taken into account. Such situation occurs in many substations in actual PSs. In [4], the maximal numbers of messages created by considered MASs and mean times of TV processes are analyzed. This paper continues the investigations of features of MASs taken into considerations in [4]. The goal of the paper is presentation of performance characteristics of these MASs in terms of probabilistic characteristics of agent activity and created messages. One should be noticed that performance requirements have to be considered for each phase of a life cycle when a new system (i.e. MAS in the paper) is designed. Such the statement is one of the essential statements in performance engineering of software systems [5], [6]. For example, the problem of performance evaluation of MASs is considered in such papers as: [7] for the systems: ZEUS, JADE and Skeleton Agents, [8] for Aglets IBM, Concordia and Voyager.

In the paper, to enhance the development of MASs the MaSE (MultiAgent Systems Engineering) methodology [9] is utilized.

2 The Utilized Idea of Power System Topology Verification

The considered MASs perform PS TV utilizing the method described in [1]. The method from [1] assumes calculation of so-called unbalance indices on the base of measurement data of active and reactive power flows at ends of branches (power lines, transformers) and voltage magnitudes at (electrical) nodes of PS. The mentioned unbalance indices are inputs for Artificial Neural Networks (ANNs) with radial basis functions. Analyzing the outputs of ANNs, decisions on existence of topology errors are taken.

For each node of PS one ANN is created. Let us that the node *i* is taken into account. Unbalance indices being inputs for the ANN associated with the node *i* are calculated using measurement data of: (i) active and reactive power flows at ends of branches which are incident to the node *i*, (ii) the voltage magnitude at the node *i*, (iii) active and reactive power flows at ends of branches which are incident to nodes neighbouring to the node *i*. The outputs of ANN for the node *i* are a base for taking decisions on correctness of modelling of branches incident to this node. One can note that there are two ANNs which take decisions about correctness of modelling the branch being between the nodes with which these ANNs are associated. The final decision is produced on the base of the mentioned decisions.

The earlier-described idea allows performing TV of a whole PS as many verifications of modelling particular branches. It ought to be underlined that results of these verifications should be coordinated each other. In the presented situation, MAS for PS TV is proposed in [2]. In [2], the agents of two types are defined. The agents of one type take decisions on the base of outputs of ANNs. The agents of second type take decisions on the base of decisions produced by the agents of the fist type for the particular branches.

3 Characteristics of the Considered Multiagent Systems

In the paper, as in [4], two MASs are considered. They are called MAS-1 and MAS-2. Their definitions are the same as in [4].

In MAS-1, there are distinguished nodal agents and the agent called *Dispatcher*. In MAS-2, apart from the same agents as in MAS-1 there are also so-called substation agents.

The nodal agent is associated with one electrical node of PS. This agent gathers measurement data of: (i) active and reactive power flows at ends of branches, which are incident to its node, (ii) the voltage magnitude at its node. The nodal agent investigates possibility of existing symptoms of occurrence of TE (TE symptoms).

At least one of the following events is considered as a TE symptom:

$$W_{Pi} \notin \left[-\delta_{WPi}, \delta_{WPi}\right], \; W_{Qi} \notin \left[-\delta_{WQi}, \delta_{WQi}\right] \qquad (1)$$

where: W_{Pi}, W_{Qi} - unbalance indices for the i-th node for active and reactive power, respectively [1]; δ_{WPi}, δ_{WQi} - positive constants.

If there are such symptoms, the nodal agent initiates LTV and in effect it takes decisions regarding correctness of modelling particular branches of PS in the area which is controlled by it. The nodal agent prepares a message for *Dispatcher*. The message contains decisions taken in the LTV process.

The substation agent is associated not with one electrical node of PS but with one substation in which there are many nodes. It performs the same functions as the nodal agent but for every node of its substation. It also takes final decisions about correctness of modelling the branches inside its substation, i.e. between electrical nodes in its substation. The substation agent prepares a message for the agent *Dispatcher*. The message contains: (i) decisions taken in the LTV processes regarding correctness of modelling the branches between a considered substation and its neighbouring ones. (ii) final decisions regarding correctness of modelling the branches inside the substation.

The agent *Dispatcher* gathers messages from the nodal and substation agents. For branches, for which the agent *Dispatcher* does not received final decisions about correctness of their modelling (from substation agents), it takes such decisions on the base of the decisions from LTVs.

4 The Analysis Model of the Multiagent System for PS TV

The analysis model of the considered MASs is built by using the AgentTool_1.8.3 tool of the MaSE technology [9]. In this model, one distinguishes goals, roles and tasks.

The main goal of the analyzed MASs is PS TV (Fig. 1). The subgoals of MASs are: management of TV and agents associated with nodes and substations of PS (the goal 1.1), executing the process of TV (the goal 1.2) and executing the process of measuring distinguished quantities in PS (the goal 1.3).

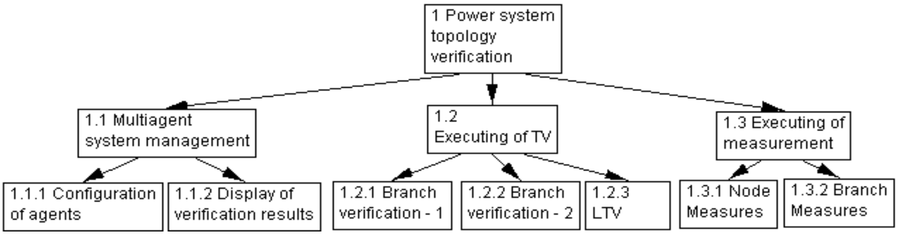

1 Power system topology verification
1.1 Multiagent system management
1.2 Executing of TV
1.3 Executing of measurement
1.1.1 Configuration of agents
1.1.2 Display of verification results
1.2.1 Branch verification - 1
1.2.2 Branch verification - 2
1.2.3 LTV
1.3.1 Node Measures
1.3.2 Branch Measures

Fig. 1. Goals diagram of MAS for PS TV

The considered systems include one agent which plays the *Dispatcher* role (a rectangle). Each of other agents plays one instance of the roles *Node* or together *Node* and *Substation* (Fig. 2b, Fig. 3b). The *Node* role (Fig. 2b, Fig. 3b) fulfils the following subgoals: 1.2.3, 1.3, 1.3.1, 1.3.2 (Fig. 1). For the *SubStation* role there is only one subgoal 1.2.2 (Fig. 3b). Other goals (Fig. 1) are secured by the *Dispatcher* role. Each role performs a few tasks (ellipses). Each task is modelled as the statecharts diagram. Using tasks, the agents of roles exchange messages with each other according to the suitable external protocols (solid lines). Internal protocols (dashed lines) are used when tasks of the same role exchange messages. MAS-1 includes two types of agents: the first one playing the *Dispatcher* role and the second one such as the nodal agent playing the *Node* role for each node of PS (Fig. 4). Additionally, MAS-2 contains the substation agents, playing the *Substation* and the *Node* roles (Fig. 5). In the last case, the *Node* roles represent particular nodes of the multi-node substations of PS.

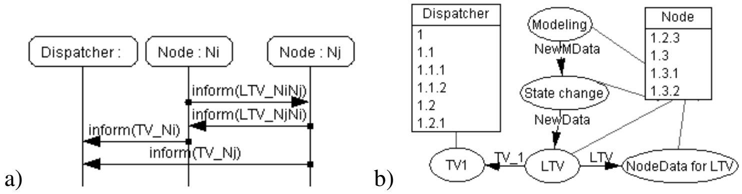


Fig. 2. The diagrams of the experimental MAS-1 for realization of the TV process: a) sequence diagram, b) role diagram

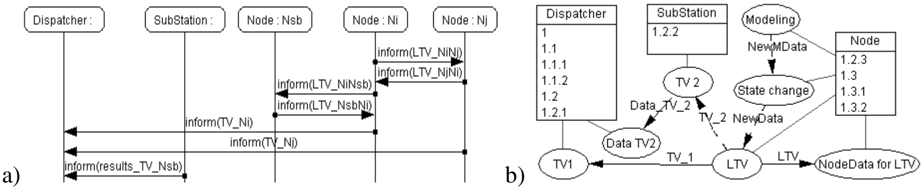


Fig. 3. The diagrams of the experimental MAS-2 for realization of the TV process: a) sequence diagram, b) role diagram

The sequences of messages exchanged by tasks are used to fulfill goals (Fig. 1) of roles (Fig. 2b, Fig. 3b). Fig. 2a presents such sequences of one instance of the *Dispatcher* role and two instances of the *Node* role for MAS-1. Additionally, Fig. 3a includes one instance of the *Substation* role and one instance of the *Node* role, both related to the power substation, modelled in MAS-2. The labels of arrows represent the messages (Fig. 2a, Fig. 3a) exchanged between the tasks of the roles with the use of the external protocols (Fig. 2b, Fig. 3b).

Further, it is assumed that the considered PS has *n* nodes, *m* branches, and the number of branches, which are connected with the considered node, is *k*. Additionally, from the view point of MAS-2 the branches of two kinds are distinguished. The branches of the first kind connect nodes belonging to different agents playing the *Substation* or *Node* roles. Let's assume that the number of such branches is equal to *p*. The branches of the second kind connect nodes belonging to the agents of the *Substation* role and they connect nodes inside the multi-node substation. The number of such branches is equal to *q*. There is $n = p + q$. The number of agents is equal to *d.* In particular, in MAS-1 where there are only agents of the *Node* roles, $d = n$.

In MAS-1, when any TE symptom is detected (Section 3) the agents of the *Node* role take independent decisions regarding correctness of modelling of branches in the *LTV* task. In MAS-2, there is the substation agent which fulfils the *Node* and *Substation* roles. The *Substation* role executes complete TV for branches, which are inside one substation.

The *Modeling* task build a PS topology model (Fig. 2b, Fig. 3b). This task internally sends its data to the *State change* task using the internal *NewMData* protocol. The *State change* task performs the detection of TE symptoms. This detection is based on the testing the nodal unbalance indices, which are earlier calculated using measurement data of active and reactive power flows at the ends of branches. The *State change* task internally sends its data to the *LTV* task for LTV by the internal *NewData* protocol.

The *LTV* task uses the external *LTV* protocol (Fig. 2b, Fig. 3b). During LTV, the total number of messages exchanged among each *LTV* task and the *Node for LTV* tasks of neighbouring nodes belonging to the different agents (the nodal agents or also the substation agents) is equal to *4p* or *4*(*m* - *q*) (each message of the *r*-units size). In other words, this number is equal to the sum of the numbers of: (i) all the *inform(LTV_NiNj)* and *inform(LTV_NjNi)* messages (Fig. 2a) exchanged among different agents of the *Node* role in the case of MAS-1 (Fig. 4); (ii) all the *inform(LTV_NiNj)*, *inform(LTV_NjNi)*, *inform(LTV_NiNsb)* and *inform(LTV_NsbNi)* messages (Fig. 3a) exchanged among agents of the *Node* roles and agents of the *Node* and *Substation* roles in the case of MAS-2 (Fig. 5). It is assumed that instances of roles *Ni* and *Nj* represent the nodal agents, whereas instances of role *Nsb is* fulfilled by substation agents. The total number of sent messages is 4*m,* in the case of MAS-1 and *4p* or 4*(m-q)* in the case of MAS-2.

The TV process is carried out after the appropriate LTV processes finish. In MAS-1, the *LTV* tasks (Fig. 2b) send at most *kn* messages (of the *r*-units size). The number *kn* is equal to 2*m*. The considered messages are the *inform(TV_Ni), inform(TV_Nj)* messages (Fig. 2a) for the *TV1* task of the agent of the *Dispatcher* role (Fig. 2b). If only some of agents of the *Node* roles detect TE symptoms, they send the *inform(TV_Ni)* messages to the agent of the *Dispatcher* role. The size of each such a message is equal to the number *kr*. The maximal number of the *inform(TV_Ni)* and *inform(TV_Nj)* messages is equal to the number of branches connected with nodes with TE symptoms.

In the case of MAS-2 as it is in the case of MAS-1, the nodal agents send messages with LTV decisions to the agent of the *Dispatcher* role. More complex situation is with respect to the substation agents. The complete TV process for the set of *q* branches, which are inside substations, is carried out by the *TV2* task of the *Substation* role of the substation agents. The final TV decisions for the mentioned branches are taken by these agents instead of the agent of the *Dispatcher* role without exchanging any messages between nodes outside the substation. Then, the *TV2* task sends the *inform(results_TV_Nsb)* message (Fig. 3a) with a final TV decision to the *Data TV2* task of the *Dispatcher* agent (Fig. 3b). The substation agent does not take final decisions for the branches which go out from its substation. For such branches only LTV decisions are taken and messages with these decisions are sent to the *TV1* task of the *Dispatcher* agent (Fig. 3b). The total number of messages sent by the nodal and substation agents to the *Dispatcher* agent is $2p+q$ or $2m-q$.

Summarizing, when all nodes identify TE symptoms, the total number of messages in the complete TV process is $6m$ for MAS-1 and $6p+q$ or $6m-5q$ for MAS-2. The mentioned numbers of messages are maximal ones.

The result from the presented analysis is that for MAS-2 the maximal number of messages exchanged among agents is smaller than it is for MAS-1. Measure of the decreasing of the mentioned number of messages is the number $5q$, where q is the number of internal branches of multi-node substations.

5 Design of Multiagent Systems for PS TV

After the analysis models are worked out, the design model of MAS-1 (Fig. 4) and MAS-2 (Fig.5) are created [9] as the *Agent Template Diagrams*. *Agent Template Diagrams* are mapped from their analysis models. These diagrams show the *Agent Communication Language* (ACL) messages [10] exchanged between agents. In Fig. 4, each instance of the *Node_Agent1* and *Node_Agen2* agents represents the nodes of PS which are connected with each other. In Fig. 5 the instances of the *SubS_Agent1* and *SubS_Agent2* agents represent substations of PS. In both MASs, the *Dispatcher* agent manages the whole PS TV.

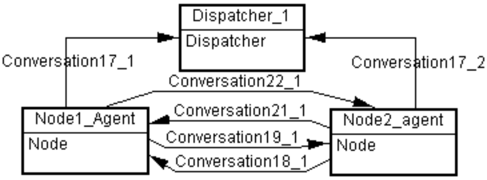


Fig. 4. The *Agent Template Diagram* for MAS-1

In MAS-1 and MAS-2, LTV is performed, whenever TE symptoms are detected. Independent decisions of the *Node_Agent1* and *Node_Agent2* agents initiate LTVs and they send the results of LTVs to the *Dispatcher* agent which executes TV for appropriate branches. Additionally, in MAS-2 the *SubS_Agent1* and *SubS_Agent2* agents execute TV (Section 4) for branches, which are internal for substations represented by the considered agents, and the final TV results are sent to the *Dispatcher* agent.

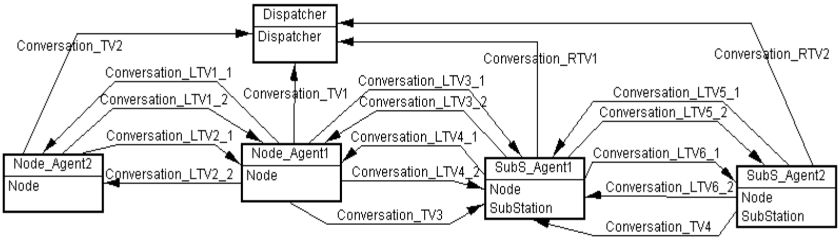


Fig. 5. The Agent Template Diagram for MAS-2

6 Description of Carried Out Investigations

During the investigations, for both the considered MASs (i.e. MAS-1 and MAS-2), probabilistic characteristics of agent activity and created messages have been determined. The mentioned characteristics are: (i) the probability of detecting at least one TE symptom, (ii) the mean number of nodes detecting a TE symptom and their neighbours for one TV cycle, (iii) mean number of messages sent during one TV cycle. The performance experiments have been done, using the modified IEEE 14-bus test system (as in [4]). The basic form of the IEEE 14-bus test system is shown in the paper [1]. The parameters of the test system are as follows: $n = 19$, $m = 25$, $p = 16$, $q = 9$. It was assumed that: (i) all branches of the test system are actually in operation, (ii) all measurement data (of active and reactive powers at ends of branches and voltage magnitudes at nodes) are burdened with Gaussian noise with the zero mean and the standard deviation determined as follows: $\sigma_P = 1/3\left(0.0035\,FS + 0.02\,M\right)$, $\sigma_Q = 1/3\left(0.006\,FS + 0.02\,M\right)$ $\sigma_V = 1/3\left[0.003(FS + M)\right]$ for active power, reactive power, and voltage magnitudes, respectively, where *FS* is a measurement range, *M* is a measured value [11], [12]. For the aim of the investigations *M* = *FS* - this means that the mentioned standard deviations have the largest possible values what corresponds with the worst possible case.

The constants in the formulae (1) are calculated as follows: $\delta_{WPi} = b\sqrt{\sum_{j\in I_i}\sigma_{Pij}^2}$, $\delta_{WQi} = b\sqrt{\sum_{j\in I_i}\sigma_{Qij}^2}$, where σ_{Pij}, σ_{Qij} are the standard deviations of small errors burdening the measurement data of active and reactive power flows, respectively, on the branch connecting the nodes *i* and *j* at the node *i; I_i* is the set of nodes connected with the node *i*; *b* is the constant equal to 1, 2 or 3.

Carrying out the investigations, it was also taken into account that errors in the model of a PS topology can exists. The assumption was made that probability of the improper modelling of a branch is p_M. p_M is the constant equal to 0.01, 0.001 or 0.0001.

The results of the investigations are in Table 1-3. The second subscript of the quantities presented in Table 1-3 points out the value of b, for which these quantities are determined.

In Table 1, the probability p_e, i.e. the probability of detecting at least one TE symptom, is shown for different p_M. The probability of detecting at least one TE symptom is defined as: $p_e = s_g / s_t$, where: s_g is a number of TV cycles for which at least one TE symptom is detected; s_t is a number of all the considered TV cycles (for which possibility of occurrence of one of the events (1) is tested). When $b = 1$, the probability of detecting at least one TE symptom in one TV cycle is very close to 1. This probability is much smaller for $b = 3$.

In Table 1, also the probability of occurrence of at least one TE (p_{TE}) is presented for different p_M. Differences between values of p_e and p_M mean that there can be cases when a TE symptom is detected but none of TEs occurs. From the view point of "a false alarm" the most favourable case is when $b = 3$.

Table 2 contains the mean numbers of nodes which detect TE symptoms ($\tilde{e}_s$) and the mean numbers of their neighbours ($\tilde{e}_n$) which do not detect TE symptoms but perform the LTV processes. Table 2 shows that there exists strong dependence of $\tilde{e}_s$ and $\tilde{e}_n$ on the parameter b. One can also note that the relation between $\tilde{e}_s$ and $\tilde{e}_n$ is different for $b = 1$ and for $b > 1$. For $b = 1$ $\tilde{e}_s > \tilde{e}_n$, but for $b > 1$ $\tilde{e}_s < \tilde{e}_n$.

The mean numbers of messages sent during the TV process are shown in Table 3. The number of messages sent during the TV process can be expressed as $M_e = M_{en} + M_{el} + M_{ef}$. For MAS-1 M_{en} is a number of the messages *inform(LTV_M_NiNj)* and *inform(LTV_M_NjNi)*. For MAS-2 M_{en} is a number of such the messages as in the case of MAS-1 and also the messages *inform(LTV_M_NiNsb)* and *inform(LTV_M_NsbNi)* (Section 4). M_{en} is determined as $M_{en} = \text{card}(\bigcup_{i=1}^{e} Z_i)$, where e is a number of nodes detecting the TE symptoms; Z_i is a set of messages sent among the i-th node, where the TE symptom is detected, and all nodes from a set $S_{a,i}$ and among nodes from the set $S_{a,i}$ and nodes from a set $S_{aa,i}$. The sets $S_{a,i}$, $S_{aa,i}$ are sets of nodes which are directly connected with the i-th node and nodes from the set $S_{a,i}$, respectively. M_{el} and M_{ef} are numbers of the messages *inform(TV_Ni)* and *inform(results_TV_Nsb)*, respectively. For MAS-1 $M_{ef} = 0$.

If the i-th node detects a TE symptom, this node and all of its neighbouring ones (from the set $S_{a,i}$) execute LTVs. The node does not execute LTV if it and none of the nodes from $S_{a,i}$ has not detected earlier a TE symptom. The node, which does not execute the LTV, sends a message to the node from $S_{a,i}$, if the last one must execute the LTV.

Further, the mean value of M_e is denoted as $\overline{M}_e$ and the mean value of M_e under the condition that there is at least one TE symptom - $\tilde{M}_e$.

Values of $\overline{M}_e$ as well as values of $\tilde{M}_e$ strongly depend on the parameter b. The smallest values of $\overline{M}_e$ and also $\tilde{M}_e$ are for $b = 3$.

For $p_M < 0.0001$, changes of values of all considered quantities are very small. Practically, these changes can be neglected.

Table 1. The probability of occurrence of at least one TE (p_{TE}) and the probability of detecting at least one TE symptom (p_{ei}) in the considered test system for different values of p_M.

p_M	p_{TE}	p_{e1}	p_{e2}	p_{e3}
0.01	0.222	1.000	0.859	0.255
0.001	0.025	1.000	0.833	0.115
0.0001	0.002	1.000	0.830	0.100

Table 2. The mean number of nodes detecting a TE symptom and their activated neighbours in one TV cycle for different values of p_M.

p_M	$\tilde{e}_{s1}$	$\tilde{e}_{s2}$	$\tilde{e}_{s3}$	$\tilde{e}_{n1}$	$\tilde{e}_{n2}$	$\tilde{e}_{n3}$
0.01	10.23	2.17	1.14	7.44	4.71	2.93
0.001	10.15	2.05	1.06	7.49	4.52	2.75
0.0001	10.14	2.04	1.05	7.49	4.50	2.74

Table 3. The mean numbers of messages sent during one TV cycle for different values of p_M in percents of the maximal numbers of these messages ($M_{e\ \max}$). For MAS-1 $M_{e\ \max} = 150$, for MAS-2 $M_{e\ \max} = 105$.

p_M	MAS-1			MAS-2			MAS-1			MAS-2		
	$\overline{M}_{e1}$	$\overline{M}_{e2}$	$\overline{M}_{e3}$	$\overline{M}_{e1}$	$\overline{M}_{e2}$	$\overline{M}_{e3}$	$\tilde{M}_{e1}$	$\tilde{M}_{e2}$	$\tilde{M}_{e3}$	$\tilde{M}_{e1}$	$\tilde{M}_{e2}$	$\tilde{M}_{e3}$
0.01	96.62	38.89	7.19	95.54	35.27	6.34	96.62	45.26	28.26	95.54	41.04	24.93
0.001	96.49	36.21	3.05	95.38	32.75	2.69	96.50	43.49	26.62	95.38	39.33	23.42
0.0001	96.48	35.93	2.63	95.36	32.50	2.30	96.48	43.31	26.45	95.36	39.16	23.27

7 Conclusion

In the paper, two MASs, i.e. MAS-1 and MAS-2 are considered. Apart from the same agents as in MAS-1, MAS-2 assumes utilization of so-called substation agents. Such solution results in improving features of MAS for PS TV. Such parameters as the maximal number of messages sent during one TV cycle $M_{e\_\max}$, the mean number of these messages for all considered TV cycle ($\overline{M}_{eb}$ $b \in \{1, 2, 3\}$) as well as the mean number of the mentioned messages for these TV cycle, in which at least one TE symptom is detected ($\tilde{M}_{eb}$ $b \in \{1, 2, 3\}$), are lower for MAS-2. The pointed out differences are relatively large. The listed parameters for MAS-2 are not larger than 70 % of the appropriate parameters for MAS-1. The considered differences are especially large when values of $\overline{M}_{e3}$ or $\tilde{M}_{e3}$ are taken into account. Additionally, it should be stressed that when the parameter b increases the here-considered parameters decreases faster for MAS-2 than for MAS-1.

For MAS-2, i.e. for the more advantageous MAS from among the considered ones, the mean number $\tilde{M}_{eb}$ $b \in \{1, 2, 3\}$, is not larger than 25 % of $M_{e\,\max}$, and the number $\overline{M}_{eb}$ $b \in \{1, 2, 3\}$ is not larger than 6.5 % of $M_{e\,\max}$. The last mean number depends strongly on the probability of occurrence of TE (p_M). It can be also stated that this last mean number is relatively low.

Generally, the results of the carried out investigations show that MAS-2 allows for specific properties of PS better than MAS-1 and in effect it enables considerable reduction of unnecessary transfer of messages in the computer network. It should be also stressed, that in this paper the impact of the operation of a computer network on the features of MASs for power system topology verification is not considered.

References

1. Lukomski, R., Wilkosz, K.: Method for Power System Topology Verification with Use of Radial Basis Function Networks. In: Sandoval, F., Prieto, A.G., Cabestany, J., Graña, M. (eds.) IWANN 2007. LNCS, vol. 4507, pp. 862–869. Springer, Heidelberg (2007)
2. Wilkosz, K.: A Multi-Agent System Approach to Power System Topology Verification. In: Yin, H., Tino, P., Corchado, E., Byrne, W., Yao, X. (eds.) IDEAL 2007. LNCS, vol. 4881, pp. 970–979. Springer, Heidelberg (2007)
3. Wilkosz, K., Kruczkiewicz, Z., Rojek, T.: Multiagent Systems for Power System Topology Verification. In: Corchado, E., Yin, H. (eds.) IDEAL 2009. LNCS, vol. 5788, pp. 815–822. Springer, Heidelberg (2009)
4. Wilkosz, K., Kruczkiewicz, Z.: Multiagent-System Oriented Models for Efficient Power System Topology Verification. In: Nguyen, N.T., Kim, C.-G., Janiak, A. (eds.) ACIIDS 2011, Part I. LNCS (LNAI), vol. 6591, pp. 486–495. Springer, Heidelberg (2011)
5. Smith, C.U., Lloyd, G.W.: Performance Solutions, A Practical Guide to Creating Responsive, Scalable Software. Addison - Wesley, Canada (2002)
6. Babczyński, T., Kruczkiewicz, Z., Magott, J.: Performance Analysis of Multiagent Industrial System. In: Klusch, M., Ossowski, S., Kashyap, V., Unland, R. (eds.) CIA 2004. LNCS (LNAI), vol. 3191, pp. 242–256. Springer, Heidelberg (2004)
7. Camacho, D., Aler, R., Castro, C., Molina, J.M.: Performance Evaluation of ZEUS, JADE, and Skeleton Agent Frameworks. In: IEEE International Conference on Systems, Man, and Cybernetics, vol. 4, p. 6 (2002)
8. Dikaiakos, M.D., Kyriakou, M., Samaras, G.: Performance Evaluation of Mobile-Agent Middleware: A Hierarchical Approach. In: Picco, G.P. (ed.) MA 2001. LNCS, vol. 2240, pp. 244–259. Springer, Heidelberg (2001)
9. Deloach, S.A.: The MaSE Methodology. In: Bergenti, F., Gleizes, M.-P., Zambonelli, F. (eds.) Methodologies and Software Engineering for Agent Systems. The Agent-Oriented Software Engineering Handbook Series: Multiagent Systems, Artificial Societes, and Simulated Organizations, vol. 11. Kluwer Academic Publishing, Dordrecht (2004)
10. Specification of FIPA, `http://www.fipa.org/specs/`
11. Dopazo, J.F., Klitin, O.A., Stagg, G.W., Van Slyck, L.S.: State Calculation of Power Systems From Line Flow Measurements. IEEE Trans. on PAS PAS-89(7), 1698–1708 (1970)
12. Dopazo, J.F., Klitin, O.A., Van Slyck, L.S.: State Calculation of Power Systems from Line Flow Measurements, Part II. IEEE Trans. on PAS PAS-91(1), 145–151 (1972)

Building a Model of an Intelligent Multi-Agent System Based on Distributed Knowledge Bases for Solving Problems Automatically

Nguyen Tran Minh Khue and Nhon Van Do

University of Information Technology
Km 20, Hanoi highway, Thu Duc District, Ho Chi Minh City, Vietnam
{khuentm,nhondv}@uit.edu.vn

Abstract. In this paper, we propose a model of an Intelligent Multi-Agent System based on three distributed knowledge bases for solving problems automatically. Besides, we present architectures of agents in the system. We also illustrate an application of this model in three fields: plane geometry, 2D analytic geometry, and algebra. In our application, we use JADE platform, Maple, MathML, XML, … Finally, we show a method to test effects of the system developed from the model of MAS proposed.

Keywords: Multi-Agent System (MAS), Intelligent Agent, Mobile Agent, problem solving, distributed knowledge bases.

1 Introduction

Up to now, models and systems about Multi-Agent have not been applied to solve problems relating to many fields about Mathematics because knowledge bases in these models and systems are not enough complex to solve them. Besides, architectures of agents and activity models of these systems are not appropriate to dealing with issues about querying knowledge or solving problems. Example, some Multi-Agent systems (MAS) such as in [1- 4], [8 -14] focused on methods to enhance effects of E-Learning. In [1], authors present a set of Technology enhanced Learning environments integrating the need aspects in a synergetic way to fit well PBLs (Project Based Learning), especially in their context of use. In [2], the intelligent tutoring module is addressed, and more specifically, its BDI (Believes, Desires, and Intentions) based agents. These agents recognize each student and obtain information about her/ his progress. So, the module suggests to each student specific tasks to achieve her/his particular learning objectives. A performance oriented approach is presented in [3]. In this approach, a set of key performance indicators (KPIs) has been set up to represent a set of measures focusing on different aspects of organizational and individual performance that are critical for the success of the organization. In [4], authors propose an approach for integrating distance learning system and knowledge management processes using knowledge creation techniques. Based on the proposed approach, an integrative framework is presented for building knowledge-based

J.-S. Pan, S.-M. Chen, N.T. Nguyen (Eds.): ACIIDS 2012, Part I, LNAI 7196, pp. 21–32, 2012.

distance learning system using the Intelligent Agent technology. In [8], a model of E-Learning was proposed based on a process of coupling of ontologies and Multi-Agent Systems for a synergy of their strengths. This model allows human agents to cooperate with software agents to automatically build courses guided by relevant learning objectives. The roles of intelligent agents within an E-Learning system, called E-University, are presented in [9]. The agents perform specific tasks on the behalf of students, professors, administrators, and other members of the university. Also a group of intelligent agents for learning activities such as user interface agents, task agents, knowledge agents and mobile agents is developed. Using the multi-agent technology in E-Learning system, gives user interaction facility to both users and designers, and adds ability to exchange information between different objects in a flexible way.

In [10], a Multi-Agent model is built to improve the effect of teaching method based on WBT (Web Based Tutoring). However, the set of rules in its Knowledge Library is not suitable to many fields. Agent based Collaborative Affective e-Learning Framework in [11] only supports for ordinary communication in virtual environment, it doesn't support for solving knowledge. The aim of this framework is to understand e-learner's affective states. In [12], the purpose of Multi-Agent e-learning system is analyzing student's errors. Based on the results of comparing student's ontology and teacher's ontology, the system generates useful advices to students. Ontology (O) includes of the set <T, R, F>. T is Term, R is Relation and F is Function. This ontology model is fairly simple; it can not deal with complex problems in reality. To build the Mobile-Agent model in [13], authors use distributed databases. Therefore, this model only can deal with issues relating titles of courses and keywords of subjects; it can not deal with contents of knowledge. In [14], the model of ontology based on hierarchy tree, so this system also can not support for solving problems automatically.

In our previous research such as in [6], we concentrated to build a model of Multi-Agent in E-Learning, and in [7] we built a language to query knowledge. In this paper, we focus on building a model of Multi-Agent for solving problems automatically. Knowledge to find answers of questions is related to three fields, and is distributed in three different places. This model simulates a complex, nature, and reasonable organization of human. In the model, we use the Computational Objects Knowledge Base (COKB) as in [5] because it is developed to support for solving problems. COKB contains 6 components: (C, H, R, Ops, Funs, Rules). C is a set of concepts, H is a set of hierarchies of concepts, R is a set of relations between concepts, Ops is a set of operations, Funs is a set of functions, Rules is a set of rules. Main issues of Multi-Agent systems include interacting between Agents, disintegrating the work, separating the tasks of Agents, and synthesizing results. These Agents act autonomously in the environment to reach their goals. They are able to recognize the state of environment, properly act and impact on environment. They coordinate together during the process of solving problems.

Some authors mentioned clustering algorithms which give the distributed systems higher performance than previous clustering algorithms. In [15], an improved algorithm based on the weighted clustering algorithm is proposed with additional constraints for selection of cluster heads in mobile wireless sensor networks. In [16], authors applied the optimization technique of genetic algorithm (GA) to the new adaptive cluster validity index, which is called Gene Index (GI). These algorithms can be applied in our system to cluster agents based on their functions and fields of knowledge. However, in this paper, we focus on the model of the system, architectures of agents, and a method to test effects of the system.

2 A Model of an Intelligent Multi-Agent System Proposed

Model of proposed MAS is shown in Fig 1. Agents reside and are executed in *Places*. A Place consists of Agents, Knowledge Base and Storage. There are four kinds of Agents : *User Agent, Intermediary Agent, Knowledge Agent* and *Notice Agent*. Among these kinds, User Agent and Knowledge Agent are kinds of *Intelligent Agents*, other Agents are kinds of *Mobile Agents*. *Knowledge Base* stores the knowledge relating to a field. *Storage* stores facts about states of local environment in the Place and facts about states of global environment in the system. In the model, we organize four Places: Place 1, Place 2, Place 3 and Place 4. *Place 1* is used to interact with users. It consists of UAgent, IAgent, IBAgent, ICAgent and Storage. UAgent is a User Agent. IAAgent, IBAgent and ICAgent are Intermediary Agents. Storage stores states of local environment of Place 1. *Place 2* is used to deal with the knowledge in field A. It consists of IAAgent, AAgent, NAAgent, Knowledge Base A, and Storage. IAAgent is an Intermediary Agent. AAgent is a Knowledge Agent in field A. NAgent is a Notice Agent of Place 2. Knowledge Base A stores the knowledge in field A. Storage stores states of local environment in Place 2 and states of global environment in the system. *Place 3* and *Place 4* are similar to Place 2, but they are for knowledge in field B or knowledge in field C. Agents communicate by ACL (Agent Communication Language). Three knowledge bases of Place 2, Place 3, and Place 4 are COKB. *Maple* is a Mathematics soft-ware integrated into the system. It supports Knowledge Agent for solving some simple and single functions.

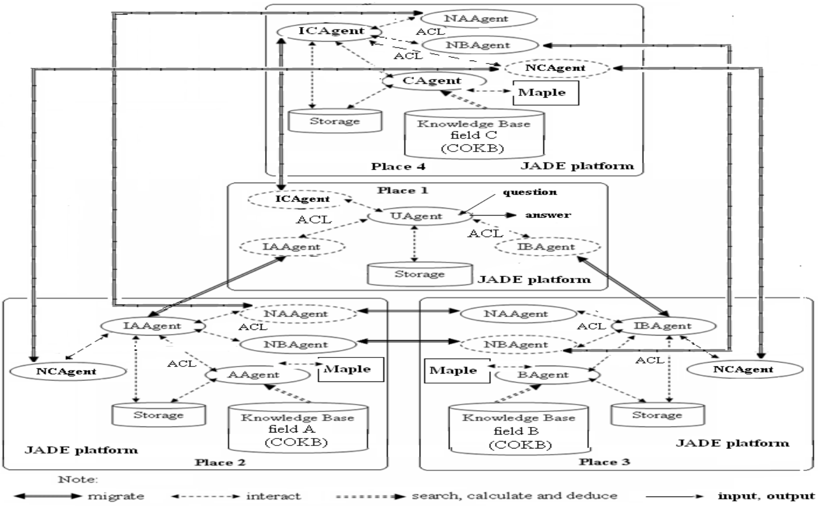


Fig. 1. A model of an Intelligent Multi-Agent system proposed

3 Activity of Intelligent Multi-Agent System Proposed

3.1 States of Local Environment in a Place

States of local environment are stored in Storages of Places. They include the first facts of questions, states about activities of Agents, the state about solving processes, facts generated after Knowledge Agents do actions. Let S_U be a set of the first facts of the question, S^0_A be a set of the first facts in field A, S^0_B be a set of the first facts in field B, S^0_C be a set of the first facts in field C. $S_U = S^0_A \bigcup S^0_B \bigcup S^0_C$

Let S_A be a set of facts about the knowledge in field A of the question, S^1_A be a set of facts generated after Knowledge Agent in field A acts. $S_A = S^0_A \bigcup S^1_A$

Generally, we have: $S_A = \bigcup_{\forall i:1 \to k} S^i_A$.

Similarly, let S_B be a set of facts about the knowledge in field B of the question. $S_B = \bigcup_{\forall j:1 \to l} S^j_B$. Let S_C be a set of facts about the knowledge in field C of the question. $S_C = \bigcup_{\forall m:1 \to n} S^m_C$

3.2 States of Global Environment in the System

States of global environment system are stored in Storages of Place 2 and Place 3. They include facts of questions, states about activities of Agents, state about solving processes in the system. Let S be a set of facts about the knowledge of the question in the system. $S = S_A \bigcup S_B \bigcup S_C$

Let G be a set of facts about goals of the problems. The state of global environment in the system 'satisfies' the goals if $S \supset G$

3.3 Rules for Agents to Choose Actions

Let F be a space of facts about the knowledge of the question. $F = \{sk_1, sk_2, \ldots., sk_n\}$.

These facts include facts about values of attributes of objects, facts about relations between objects, and facts about relations between attributes of an object. Let A be a set of actions of Agents. $A = \{a_1, a_2, \ldots, a_m\}$. These actions include searching the knowledge, calculating functions in Knowledge Base, and deducing based on rules in Knowledge Base. Let AcRules be a set of rules for Agent to choose actions according to conditional facts. A rule in AcRules has a form as following: $C \Rightarrow a$ with $C \subset F$, $a \in A$. Let S be a set of facts about the knowledge of the question in the system, $S \subset F$. Let A_S be a set of actions of Agents 'suitable' to the state of global environment S.

$$A_S = \bigcup_{(C \Rightarrow a) \text{ with } (C \subseteq S)} \{a\}$$

We have $A_S = \{a_{i, ...,} a_j\}$, a_i is an action of a Knowledge Agent. Let w_i be a priority level of acting a_i. Priority levels simulate principles to choose actions naturally of human. Knowledge Agent chooses action a_i whose priority level is highest. After Knowledge Agent acts, new facts are generated. The state of local environment changes in its Place. That makes the state of global environment in the system changes.

3.4 An Activity Model of the System

An activity model of the system is shown in Fig.2. User input the question into the system. Agents co-ordinate to find answers. User Agent disintegrates into knowledge facts of the question. These facts and other facts of the environment are the first state of global environment in the system. If the state of global environment 'satisfies' the goals, answers are found. If not, Knowledge Agents choose actions 'suitable' to the current state of global environment. If all of Knowledge Agents don't find any action, no answer is found. If at least a Knowledge Agent finds an action, Knowledge Agents do action chosen. The state of global environment is updated by Intermediary Agents and Notice Agents. The process is repeated until the state of global environment 'satisfies' the goals, then answers are found.

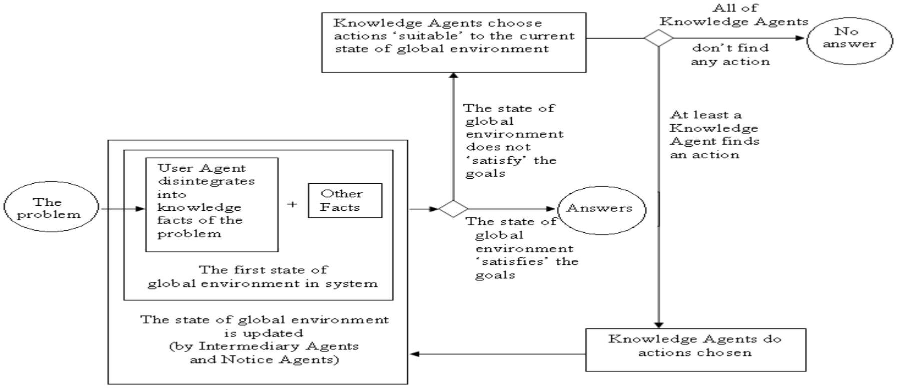

Knowledge Agents choose actions 'suitable' to the current state of global environment
All of Knowledge Agents
don't find any action
No answer
The state of global environment does not 'satisfy' the goals
At least a Knowledge Agent finds an action
The problem
User Agent disintegrates into knowledge facts of the problem
+
Other Facts
The first state of global environment in system
Answers
The state of global environment 'satisfies' the goals
The state of global environment is updated (by Intermediary Agents and Notice Agents)
Knowledge Agents do actions chosen

Fig. 2. Activity model of the system

4 Architectures of Agents

4.1 An Architecture of a User Agent

Architecture of a User Agent is shown in Fig.3. A user inputs the question/problem into the system according to given structures through the *Received Problem.* This module sends the problem to the *Disintegrate Problem.* It disintegrates the problem into objects, relations and facts. They are sent to the *Classify Knowledge* and divided into three groups: the knowledge in field A, the knowledge in field B and the knowledge in field C.

The *Knowledge of Agent* memorizes them, sends them to the *Know the state of local environment*, and then them are updated into the *Storage*. The *Knowledge of Agent* sends the knowledge of the problem to Intermediary Agents. After these Agents migrate to other Places and back with answers, it receives answers and sends them to the *Synthesize Answers*. This module synthesizes them into final answers. Then it sends them to the *Show Answers*. The *Show Answers* presents final answers to the user.

4.2 An Architecture of an Intermediary Agent

Architecture of an Intermediary Agent is shown in Fig.4. Intermediary Agent interacts with User Agent to receive facts of the problem through the *Interact User Agent*. The *Knowledge of Agent* memorizes them and activates the *Dispatch Agent to its Place*. After Agent migrates, it sends facts to Knowledge Agent. After that it receives new facts from this Agent. Then, the *Knowledge of Agent* finds available Notice Agent in its Place. If there is no Notice Agent, it activates the *Create Agent* to create a new Notice Agent. The *Knowledge of Agen*t sends Notice Agent the notice about the change of local environment. Similarly, if it receives the notice from a Notice Agent of other Place, it updates facts in the notice into the *Storage*. If it receives answers from Knowledge Agent, it activates the *Retract Agent back to previous Place*. After Agent migrates, the *Knowledge of Agent* sends answers to User Agent.

4.3 An Architecture of a Knowledge Agent

Architecture of a Knowledge Agent is shown in Fig.5. The *Knowledge of Agent* receives and memorizes the facts about the problem from Intermediary Agent. It sends them to the *Know the state of local environment*, and then they are updated into the *Storage*. After that it activates the *Check Goals*. If the state of global environment ‘satisfies’ the goals, it sends answers to Intermediary Agent. If the state doesn’t ‘satisfy’ the goals, it sends the facts to the *Check Rule*. Based on the *Rules for Agent to act*, this module chooses rules. If there is no rule appropriate to the state, no action is chosen. In this case, it sends this fact to the *Know the state of local environment* and Intermediary Agent. The fact is updated into the *Storage*. If the *Check Goal* finds appropriate rules, the *Choose Action* is activated. After that the *Do action* searches, calculates, and deduces by using the knowledge of the problem and the knowledge in Knowledge Base. After acting, it generates new facts about the problem. The *Knowledge of Agent* memorizes them and sends them to the *Know the state of local environment* and Intermediary Agent. The *Storage* is updated and the *Check Goals* is activated. The processes repeat until the state of global environment ‘satisfies’ the goals. Then, the *Knowledge of Agent* sends answers to Intermediary Agent.

4.4 An architecture of a Notice Agent

Architecture of a Notice Agent is shown in Fig.6. The *Knowledge of Agent* receives the notice from Intermediary Agent. It activates the *Dispatch Agent to other Place*. After Agent migrates, it announces the notice to Intermediary Agent of that Place.

Then, it receives the result to know if announcing the notice successful or not. After that it actives the *Retract Agent back to its Place*. After Agent migrates, it informs the result to Intermediary of its Place. The *Knowledge of Agen*t can active the *Dispose Agent* to kill Agent if Agent is lost in a long time while migrating.

5 An Application

We build an application for solving problems relating to three fields: plane geometry (field A), 2D analytic geometry (field B), and algebra (field C). Users input questions according to given structures. Agents in the system coordinate to solve the problems and shows answers to users. We used JADE 4.0, JDK 1.6, Maple 13, MathML, XML, dom4j, Suim, MozSwing, … in our application. Example question 1: In co-ordinate plane Oxy sets A[1,3], B[4,2], D $\in$ Ox, DA=DB. Find coordinate of point D and perimeter of OAB. The process of solving problem as following: The first goal of the system is to find coordinate of point D. The system searches a concept about the coordinate of a point in 2D analytic geometry Knowledge Base. Then, the system finds xD and yD. The system use knowledge about 2D analytic geometry to calculate yD . We have: D $\in$ Ox => yD = 0

In the following process, the system use knowledge about 2D analytic geometry to calculate the module of DA and DB.

DA=DB => $\sqrt{(xA-xD)^2+(yA-yD)^2}=\sqrt{(xB-xD)^2+(yB-yD)^2}$

=> $\sqrt{(1-xD)^2+(3-0)^2}=\sqrt{(4-xD)^2+(2-0)^2}$

Then, the system use knowledge about algebra to find value of xD from following equation: $\sqrt{(1-xD)^2+(3-0)^2}=\sqrt{(4-xD)^2+(2-0)^2}$

=> xD = 5/3 (use Maple to find solutions)

The system use knowledge about plane geometry to calculate perimeter of triangle OAB. Perimeter (OAB) = OA + AB + OB . After calculating modules of OA, AB, OB, the system receives results: Perimeter (OAB) = $2\sqrt{10}+2\sqrt{5}$

6 A Method to Test Effects of the System

We built two systems and a test-application. *System 1* is built to solve problems in Multi-Agent paradigm base the above model. *System 2* is built to solve problems in Client-Server paradigm. In this system, we also use the same knowledge bases and the same solving solution as system 1. Based on the results, we recognize that using Multi-Agent is better and the model proposed is reasonable. The comparing is done by test-application. In this application, we use TCL (Tool Command Language) which is a strong scripting language and an interpreter for that language. By automation testing, we can compare the performance of system 1 and system 2. We focus on testing functions of two systems when they are run in the network, whose connection is not continuous. We test two systems automatically with 2000 test cases and 100 questions. The picture of Multi-Agent Paradigm and Client-Server Paradigm is shown in Fig. 7.

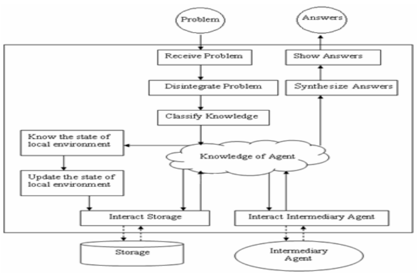


Fig. 3. Architecture of a User Agent

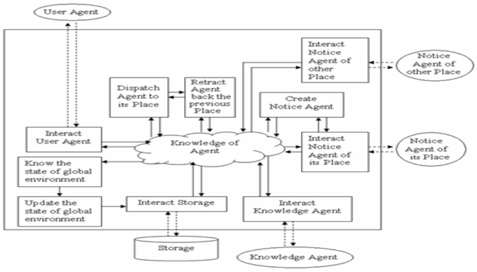

User Agent
Interact Notice Agent of other Place
Notice Agent of other Place
Dispatch Agent to its Place
Retract Agent back the previous Place
Create Notice Agent
Interact User Agent
Knowledge of Agent
Interact Notice Agent of its Place
Notice Agent of its Place
Know the state of global environment
Update the state of global environment
Interact Storage
Interact Knowledge Agent
Storage
Knowledge Agent

Fig. 4. Architecture of Intermediary Agent

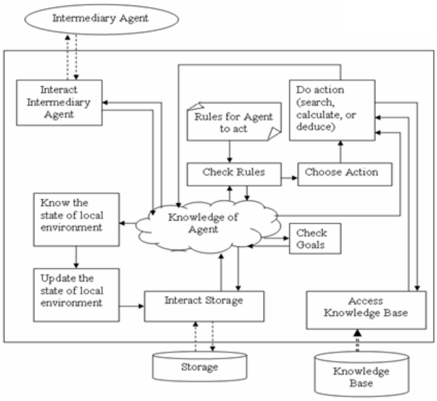


Fig. 5. Architecture of a Knowledge Agent

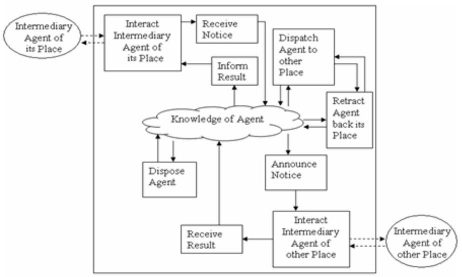

Intermediary Agent of its Place
Interact Intermediary Agent of its Place
Receive Notice
Inform Result
Dispatch Agent to other Place
Knowledge of Agent
Retract Agent back its Place
Dispose Agent
Announce Notice
Receive Result
Interact Intermediary Agent of other Place
Intermediary Agent of other Place

Fig. 6. Architecture of a Notice Agent

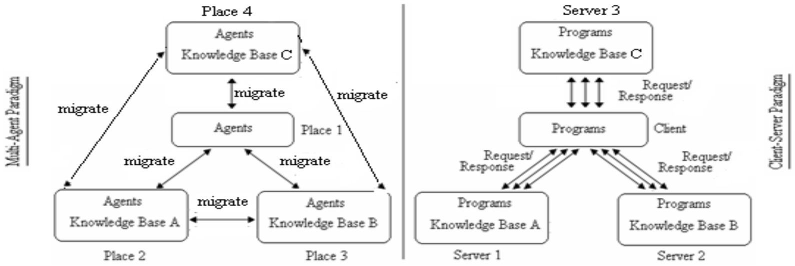


Fig. 7. Multi-Agent vs Client-Server

The model in the left part of Fig.7 is an outline of the model in Fig.1. The model in Fig 1 is detailed to show kinds of agents, relationships of agents, and activity model of the proposed system. We can compare the model in the left part of Fig.7 (Multi-Agent Paradigm) with the model in the right part of Fig 7 (Client-Server Paradigm). In Multi-Agent Paradigm, the network load is reduced. After an agent migrates to other Place, it interacts in the locality. Therefore, the intermediate calculations between Places are reduced. Agents can interact each other. The Multi-Agent paradigm is decentralized. Here are two test cases to illustrate that agents can work with the uncontinuous connection.

Test case 1: test functions of system 1 in Multi-Agent Paradigm if the connection between Place 1 and Place 3 is interrupted after IBAgent migrates from Place 1 to Place 3, using question 1. Results of Test case 1 are shown in Table 1. We see that all results of testing functions are "Passed".

Table 1. Results of Test case 1

Connection status	Function	Result
Place 1-2: OK, Place 1-3: OK, Place 1-4: OK	UAgent receives the question	Passed
Place 1-2: OK, Place 1-3: OK, Place 1-4: OK	UAgent disintegrates the question	Passed
Place 1-2: OK, Place 1-3: OK, Place 1-4: OK	UAgent classifies the question	Passed
Place 1-2: OK, Place 1-3: OK, Place 1-4: OK	UAgent interacts with IAAgent, IBAgent, ICAgent	Passed
Place 1-2: OK, Place 1-3: OK, Place 1-4: OK	IBAgent migrates from Place 1 to Place 3	Passed
Place 1-2: OK, Place 1-3: Failed, Place 1-4: OK	IBAgent sends knowledge to BAgent	Passed
Place 1-2: OK, Place 1-3: Failed, Place 1-4: OK	BAgent searches, calculates, deduces	Passed
Place 1-2: OK, Place 1-3: Failed, Place 1-4: OK	IBAgent receives new facts from BAgent	Passed
Place 1-2: OK, Place 1-3: Failed, Place 1-4: OK	IBAgent sends NBAgent the change	Passed
Place 1-2: OK, Place 1-3: Failed, Place 1-4: OK	NBAgent migrates from Place 3 to Place 2	Passed
Place 1-2: OK, Place 1-3: Failed, Place 1-4: OK	…	Passed
Place 1-2: OK, Place 1-3: Failed, Place 1-4: OK	IBAgent receives results from BAgent	Passed
Place 1-2: OK, Place 1-3: OK, Place 1-4: OK	IBAgent migrates from Place 3 to Place 1	Passed
Place 1-2: OK, Place 1-3: OK, Place 1-4: OK	…	Passed
Place 1-2: OK, Place 1-3: OK, Place 1-4: OK	UAgent shows results to the user	Passed

Test case 2: test functions of system 2 in Client-Server Paradigm if the connection between Client and Server 2 is interrupted after the question is sent from Client to Server 2, using question 1. Results of Test case 2 are shown in Table 2. We see that some results of testing functions are "Failed" and final result is also "Failed".

Table 2. Results of Test case 2

Connection Status	Function	Result
Client-Server1:OK,Client-Server2:OK,Client-Server3:OK	Program 1 at Client receives the question	Passed
Client-Server1:OK,Client-Server2:OK,Client-Server3:OK	Program 1 disintegrates the question	Passed
Client-Server1:OK,Client-Server2:OK,Client-Server3:OK	Program 1 classifies the question	Passed
Client-Server1:OK,Client-Server2:OK,Client-Server3:OK	Program 1 sends knowledge to Program 2,3,4 at Client	Passed
Client-Server1:OK,Client-Server2:OK,Client-Server3:OK	Program 3 sends knowledge to Program 5 at Server 2	Passed
Client-Server1:OK,Client-Server2:Failed,Client-Server3:OK	Program 5 sends knowledge to Program 6 at Server 2	Passed
Client-Server1:OK,Client-Server2:Failed,Client-Server3:OK	Program 6 searches, calculates, deduces	Passed
Client-Server1:OK,Client-Server2:Failed,Client-Server3:OK	Program 5 receives new facts from Program 6	Passed
Client-Server1:OK,Client-Server2:Failed,Client-Server3:OK	Program 5 sends Program 7 at Server 2 the change	Passed
Client-Server1:OK,Client-Server2:Failed,Client-Server3:OK	Program 7 sends the notice to Program 3	Failed
Client-Server1:OK,Client-Server2:Failed,Client-Server3:OK	Program 3 sends the notice to Program 7	Failed
Client-Server1:OK,Client-Server2:Failed,Client-Server3:OK	…	Failed
Client-Server1:OK,Client-Server2:Failed,Client-Server3:OK	Program 5 receives results from Program 6	Failed
Client-Server1:OK,Client-Server2:OK,Client-Server3:OK	Program 5 sends results to Program 3	Failed
Client-Server1:OK,Client-Server2:OK,Client-Server3:OK	…	Failed
Client-Server1:OK,Client-Server2:OK,Client-Server3:OK	Program 1 shows results to the user	Failed

After testing, we see that in 93% cases, the system in Multi-Agent Paradigm shows the final results to the user when it is run in the network, whose connection is not continuous. In other cases, it stops working when the time the connection is interrupted is more than the time of life-cycle of agent. In all cases, the system in Client-Server Paradigm can not show the final results to the user in the network, whose connection is not continuous.

7 Conclusion

In this paper, we present some results in building a model of an Intelligent Multi-Agent System based on distributed knowledge bases for automatically solving problems automatically. The model has some advantages. Firstly, it is open that means the model can be changed by adding or separating the knowledge and Agents as well. Secondly, it has flexibility and intelligence, because it can solve complex issues based on distributed knowledge bases that are able to be updated. Thirdly, it supports effectively for organizing distributed knowledge bases to solve complicated problems. The knowledge consists of many knowledge bases relating to each other in many fields and it is distributed in many different places. We also describe an application of this model in three fields: algebra, plane geometry and 2D analytic geometry. Specially, the Multi-Agent system has highly flexible capable to solve problems about knowledge. In the future, we will develop the model so that it can solve many problems concurrently. Besides, we will build Agents able to update knowledge bases.

References

1. El Kamoun, N., Bousmah, M., Aqqal, A., Morocco, E.J.: Virtual Environment Online for the Project based Learning session. Cyber Journals: Multidisciplinary Journals in Science and Technology, Journal of Selected Areas in Software Engineering (JSSE) January Edition (2011)
2. Mikic-Fonte, F.A., Burguillo-Rial, J.C., Llamas-Nistal, M.: A BDI-based Intelligent Tutoring Module for the e-Learning Platform INES. In: 40th ASEE/IEEE Frontiers in Education Conference, Washington, DC (2010)
3. Wang, M., Ran, W., Jia, H., Liao, J., Sugumaran, V.: Ontology-Based Intelligent Agents in Workplace eLearning. In: Americas Conference on Information Systems, AMCIS (2009)
4. Thuc, H.M., Hai, N.T., Thuy, N.T.: Knowledge Management Based Distance Learning System using Intelligent Agent. Posts, Telecommunications & Information Technology Journal, Special Issue: Research and Development on Information and Communications Technology (16), 59–69 (2006)
5. Van Nhon, D.: Construction and development models of knowledge representation for solving problems automatically, Ph.D Thesis, University of National Science, Ho Chi Minh City (2001)
6. Van Nhon, D., Khue, N.T.M.: Building a model of Multi-Agent systems and its application in E-Learning. Posts, Telecommunications & Information Technology Journal, Special Issue: Research and Development on Information and Communications Technology (18), 100–107 (2007)

7. Khue, N.T.M., Hai, N.T.: Developing a language based on FIPA-ACL to query knowledge in a multiagent system. In: Proceedings of the Sixth International Conference on Information Technology for Education and Research, Vietnam, pp. 176–183 (2010)
8. El Bouhdidi, J., Ghailani, M., Abdoun, O., Fennan, A.: A New Approach based on a Multi-ontologies and Multiagents System to Generate Customized Learning Paths in an E-Learning Platform. International Journal of Computer Applications (0975 – 8887) 12(1) (December 2010)
9. El-Bakry, H.M., Mastorakis, N.: Realization of E-University for Distance Learning. WSEAS TRANSACTIONS on Computers 8(1) (2009)
10. Yoshida, T.: Cooperation learning in Multi-Agent Systems with annotation and reward. International Journal of Knowledge-based and Intelligent Engineering System 11, 19–34 (2007)
11. Neiji, M., Ben Ammar, M.: Agent based Collaborative Affective e-Learning Framework. The Electronic Journal of E-Learning 5(2), 123–134 (2007)
12. Gladun, A., Rogustshina, J.: An ontology based approach to student skills in Multi-Agent e-Learning systems. International Journal Information Technologies and Knowledge 1 (2007)
13. Quah, J.T.S., Chen, Y.M., Leow, W.C.H.: E-Learning System, Nanyang Technological University, Team LiB (2004)
14. Tecuci, G., Boicu, M., Marcu, D., Stanescu, B., Boicu, C., Comello, J.: Training and Using Disciple Agents, US Army War College (2004)
15. Hong, T.-P., Wu, C.-H.: An Improved Weighted Clustering Algorithm for Determination of Application Nodes in Heterogeneous Sensor Networks. Journal of Information Hiding and Multimedia Signal Processing 2(2), 173–184 (2011)
16. Lin, T.C., Huang, H.C., Liao, B.Y., Pan, J.S.: An Optimized Approach on Applying Genetic Algorithm to Adaptive Cluster Validity Index. International Journal of Computer Sciences and Engineering Systems 1(4), 253–257 (2007)

Temporal Reasoning in Multi-agent Workflow Systems Based on Formal Models

Fu-Shiung Hsieh and Jim-Bon Lin

Department of Computer Science and Information Engineering
Chaoyang University of Technology
41349 Taichung, Taiwan
fshsieh@cyut.edu.tw

Abstract. A critical issue in patient planning is to determine whether the medical processes of a patient can be completed by a time constraint based on the available resources in hospitals. The problem is a Temporal Constraint Satisfaction Problem (TCSP). The objectives of this paper are to propose a viable and systematic approach to develop a distributed cooperative problem solver for TCSP and estimate the shortest and the longest completion time for handling a patient in the presence of uncertainty based on Multi-agent systems (MAS) architecture. Our approach combines MAS with a subclass of time Petri net (TPN) models to solve TCSP. Existing analysis methods of TPN based on state classes cannot be applied directly due to distributed architecture of MAS. In this paper, a temporal analysis method based on MAS architecture is proposed. Our temporal analysis method efficiently deduces the earliest and latest completion time of a patient based on interaction between agents.

Keywords: Temporal reasoning, workflow, multi-agent system.

1 Introduction

Patient planning in hospitals is a highly complex task due to distributed organizational structure, dynamic medical processes, uncertainty in operation time and stochastic arrival of urgent patients. In existing literature, there are several studies on planning and scheduling in hospitals. For example, Decker and Jinjiang propose a MAS solution using the Generalized Partial Global Planning approach that preserves the existing organization structures while providing better performance [1]. Kutanoglu and Wu investigate a new method based on combinatorial auction [2]. Oddi and Cesta explore constraint based scheduling techniques and implement a mixed-initiative problem solving approach to managing medical resources in a hospital [3]. Daknou, Zgaya, Hammadi and Hubert focus on treatment scheduling for patients at emergency department in hospitals [4] based on MAS. In hospitals, an urgent patient often needs to be handled properly by a time constraint. A critical issue is to determine whether the medical processes of a patient can completed by a time constraint in the presence of uncertainty based on the available resources in hospitals. The problem is a Temporal Constraint Satisfaction Problem (TCSP). The objective of this paper is to propose a viable approach to develop a problem solver for TCSP.

J.-S. Pan, S.-M. Chen, N.T. Nguyen (Eds.): ACIIDS 2012, Part I, LNAI 7196, pp. 33–42, 2012.

In existing AI literature, Constraint Satisfaction Problem (CSP) has been extensively studied [9]. Conry *et al.* propose a multistage negotiation paradigm for solving DCSP [17]. Although the method for DCSP can be applied in MAS, it cannot be applied directly to TCSP due to the lack of a model to specify the temporal relation of different operations in medical processes and a method for temporal reasoning. MAS [6, 7] provides a flexible architecture to dynamically allocate resources in a hospital to handle patients. Resources such as doctors, staffs, specialists and nurses can all be modeled as agents. The key issue is how to make these agents work together coherently to handle a patient timely. To develop a problem solver, the concept of cooperative distributed problem solving (CDPS) is adopted [8] to make agents work together to solve problems. To solve TCSP, we combine CDPS with a formal temporal model and propose a problem solving architecture with three types of agents, including workflow agents and resource agents and patient agents. We construct models for workflow agents and resource agents. A medical process usually involves a complex process described by a workflow. To facilitate reasoning on time, we adopt time Petri nets (TPN) [13] as the modeling tool in this paper instead of using informal workflow specification languages such as XPDL [10], BPMN [11] and WS-BPEL [12]. The static interval of a transition in TPN specifies the earliest time and the latest time for firing a transition since it is enabled. It is suitable for describing the uncertainty in time. Our TPN model is different from the one proposed by Merlin and Farbor [13] as the static interval of a transition may include nonnegative real numbers instead of nonnegative rational numbers. Existing analysis methods of TPN based on the construction of state class graphs cannot be applied directly to MAS. We propose a temporal reasoning method based on interactions between patient agents, workflow agents and resource agents using contract net protocol (CNP)[14] to efficiently deduce the completion time. We develop a problem solver based on Java Agent Development Environment (JADE) to verify our method.

The remainder of this paper is organized as follows. In Section 2, we formulate TCSP. In Section 3, we introduce TPN models for agents. We study temporal properties of the TPN models in Section 4. In Section 5, we propose an agent interaction mechanism based on CNP. We conclude this paper in Section 6.

2 Temporal Constraint Satisfaction Problem

A typical medical process may consist of a number of operations such as registration, diagnosis, radiology test, blood examination, anesthesia, surgery, intensive and discharge, etc. These operations are performed by different hospital workers such as doctor, staff, specialist and nurse. We propose a problem solving environment based on MAS to determine a patient can be handled timely based on the available resources in a hospital. There are three types of agents corresponding, including workflow agents, resource agents (e.g. doctor agents, staff agents, specialist agents and nurse agents) and patient agents. The problem is to determine whether there exists a set of workflow agents and a set of resource agents with available time slots that can coherently handle a patient by a time constraint. MAS provides a flexible architecture for agents to discover each other to form a dynamic team to handle patients. In this paper, we adopt MAS to study the problem and develop our design methodology. To state this problem, we need the following definition. Let ψ denote the time constraint for completing a patient's medical processes. Let $\boldsymbol{RA}$ denotes the set of resource agents and $\boldsymbol{WA}$ denotes the set of workflow agents. In hospitals, the processing time for operations is usually highly uncertain. The uncertainty in operation time for an operation performed by an agent in $\boldsymbol{RA}$ is specified by an interval [α ; β], where α and β are the lower bound and the upper bound on the operation time. A dynamic organization is denoted by $H(WA, RA)$, where WA denotes the set of workflow agents in H and RA denotes the set of resource agents that take part in the activities in H . The Temporal Constraint Satisfaction Problem (TCSP) is stated as follows. Given a time constraint ψ , a set of resource agents $\boldsymbol{RA}$, a set of workflow agents $\boldsymbol{WA}$ and the lower bound and the upper bound on the processing time for each operation performed by the resource agent, the problem is to determine whether there exist $RA \subseteq \boldsymbol{RA}$ and $WA \subseteq \boldsymbol{WA}$ such that the shortest completion time and the longest completion time of $H(WA, RA)$ satisfy ψ .

The Java Agent Development Environment (JADE) platform that provides a built-in directory service through DF agent (Directory Facilitator) agent to simplify service publication and discovery in the development of MAS. We develop a problem solver based on JADE as shown in Fig. 1. To study TCSP, a mathematical model for each workflow agent and resource agent is proposed in the next section.

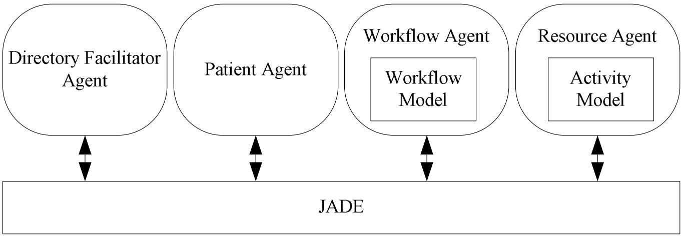


Fig. 1. Architecture to solve TCSP

3 Modeling Workflows and Activities of Agents

The medical processes in a hospital usually form complex workflows. As the processing time of each step or operation in the medical processes cannot be specified exactly a priori and only statistical data is available based on historical data, a proper model must be able to capture the uncertainty in processing time. TPN [13,15] is a model that meets these requirements as there are a lot of theories and tools that support its modeling and analysis. Therefore, we adopt TPN as our modeling tool to capture the synchronous/asynchronous and concurrent activities in hospitals. A TPN [15] G is a five-tuple $G = (P, T, F, C, m_0)$, where P is a finite set of places, T is a finite set of transitions, $F \subseteq (P \times T) \cup (T \times P)$ is the flow relation, $C : T \to R^+ \times (R^+ \cup \infty)$ is a mapping called static interval, which specifies the time interval(s) for firing each transition, where R^+ is the set of nonnegative real numbers, and $m_0 : P \to Z^{|P|}$ is the initial marking of the PN with Z as the set of nonnegative integers. A Petri net with initial marking m_0 is denoted by $G(m_0)$. A marking of G is a vector $m \in Z^{|P|}$ that indicates the number of tokens in each place under a state. $^\bullet t$ denotes the set of input places of transition t . A transition t is enabled and can be fired under m iff $m(p) \geq F(p,t) \forall p \in {}^\bullet t$. In a TPN, each transition is associated with a time interval [α ; β]; where α is called the static earliest firing time, β is called the static latest firing time, and $\alpha \leq \beta$; where α ($\alpha \geq 0$) is the minimal time that must elapse since the transition is enabled before firing the transition and β ($0 \leq \beta \leq \infty$) is the maximum time during which the transition is enabled without being fired. Firing a transition removes one token from each of its input places and adds one token to each of its output places. A marking m' is reachable from m if there exists a firing sequence s bringing m to m' . A TPN $G = (P, T, F, C, m_0)$ is live if, no matter what marking has been reached from m_0 , it is possible to ultimately fire any transition of G by progressing through some further firing sequence.

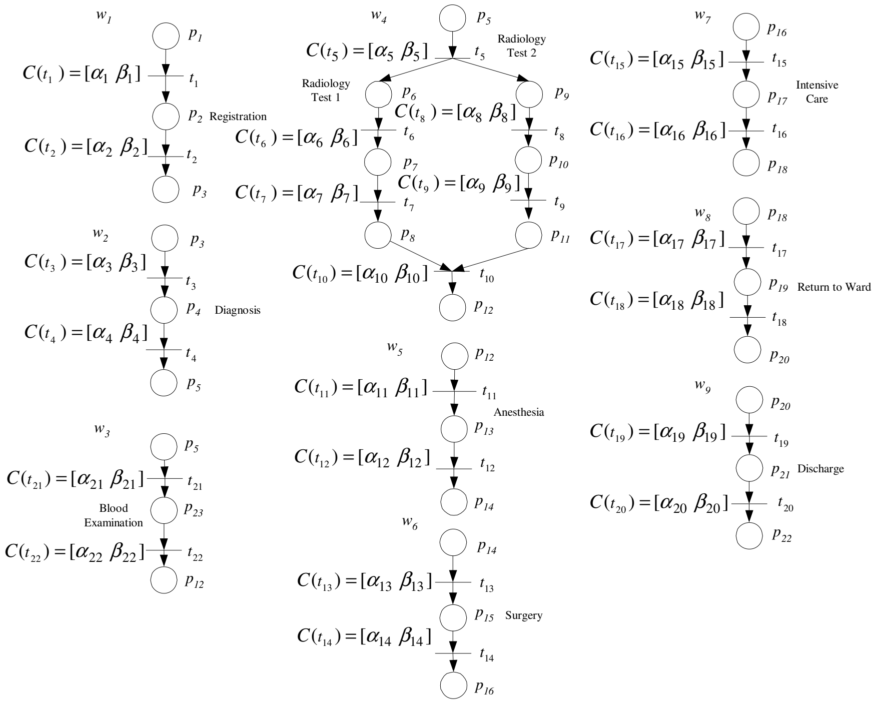

w_1
p_1
$C(t_1)=[\alpha_1\ \beta_1]$
t_1
p_2 Registration
$C(t_2)=[\alpha_2\ \beta_2]$
t_2
p_3
w_2
p_3
$C(t_3)=[\alpha_3\ \beta_3]$
t_3
p_4 Diagnosis
$C(t_4)=[\alpha_4\ \beta_4]$
t_4
p_5
w_3
p_5
$C(t_{21})=[\alpha_{21}\ \beta_{21}]$
t_{21}
Blood Examination
p_{23}
$C(t_{22})=[\alpha_{22}\ \beta_{22}]$
t_{22}
p_{12}
w_4
p_5
$C(t_5)=[\alpha_5\ \beta_5]$
t_5
Radiology Test 2
Radiology Test 1
p_6
p_9
$C(t_8)=[\alpha_8\ \beta_8]$
$C(t_6)=[\alpha_6\ \beta_6]$
t_6
t_8
p_7
p_{10}
$C(t_9)=[\alpha_9\ \beta_9]$
$C(t_7)=[\alpha_7\ \beta_7]$
t_7
t_9
p_8
p_{11}
$C(t_{10})=[\alpha_{10}\ \beta_{10}]$
t_{10}
p_{12}
w_5
p_{12}
$C(t_{11})=[\alpha_{11}\ \beta_{11}]$
t_{11}
Anesthesia
p_{13}
$C(t_{12})=[\alpha_{12}\ \beta_{12}]$
t_{12}
p_{14}
w_6
p_{14}
$C(t_{13})=[\alpha_{13}\ \beta_{13}]$
t_{13}
p_{15} Surgery
$C(t_{14})=[\alpha_{14}\ \beta_{14}]$
t_{14}
p_{16}
w_7
p_{16}
$C(t_{15})=[\alpha_{15}\ \beta_{15}]$
t_{15}
Intensive Care
p_{17}
$C(t_{16})=[\alpha_{16}\ \beta_{16}]$
t_{16}
p_{18}
w_8
p_{18}
$C(t_{17})=[\alpha_{17}\ \beta_{17}]$
t_{17}
p_{19} Return to Ward
$C(t_{18})=[\alpha_{18}\ \beta_{18}]$
t_{18}
p_{20}
w_9
p_{20}
$C(t_{19})=[\alpha_{19}\ \beta_{19}]$
t_{19}
p_{21} Discharge
$C(t_{20})=[\alpha_{20}\ \beta_{20}]$
t_{20}
p_{22}

Fig. 2. TPN models of workflow agents $w_1 \sim w_9$

To model a medical workflow as a TPN, we use a place to denote a state in the workflow while a transition denotes an operation. The workflow of a medical process w_n is modeled by extending a subclass of Petri nets defined as follows.

Definition 3.1: A connected acyclic time marked graph (CATMG) w'_n = ($P'_n, T'_n, F'_n, C'_n, m'_{n0}$) is a connected marked graph without any cycle.

Definition 3.2: A transition in T'_n without any input place is called a terminal input transition. A transition in T'_n without output place is called a terminal output transition.

In this paper, we consider the class of CATMG w'_n = ($P'_n, T'_n, F'_n, C'_n, m'_{n0}$) with a single terminal input transition and a single terminal output transition.

Definition 3.3: The workflow of a workflow agent w_n is an acyclic TPN w_n = (P_n , T_n , F_n , C_n , m_{n0}) obtained by augmenting a CATMG w'_n = ($P'_n, T'_n, F'_n, C'_n, m'_{n0}$) with a service input place ε_n and a service output place θ_n and adding an arc connecting ε_n to the terminal input transition of w'_n and adding an arc connecting the terminal output transition to θ_n .

Fig. 2 illustrates the individual workflow Petri net models w_1 through w_9 for the steps of a medical process consisting of registration, diagnosis, radiology test, blood examination, anesthesia, surgery, intensive and discharge, etc. Note that the individual workflow TPN model for each step has one start state, several processing state and one finished state. For example, the workflow Petri net model w_1 for Step 1 consists of one start state (p_1), one registration state (p_2) and one finished state (p_3).

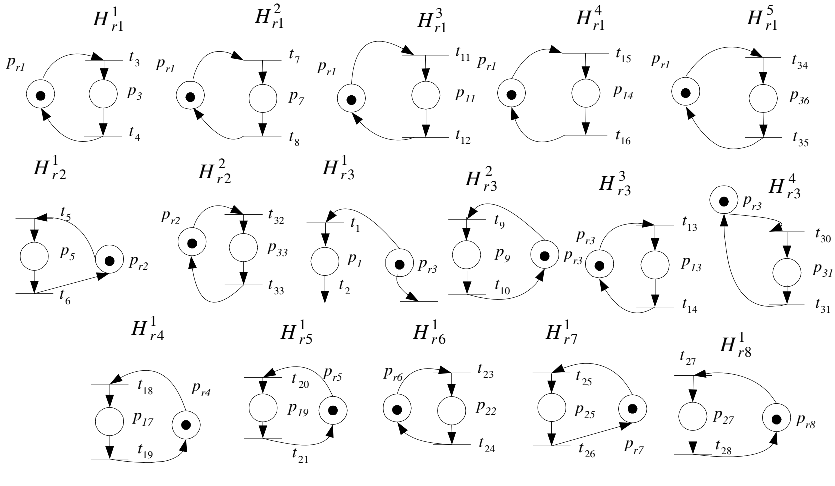


Fig. 3. Resource Activities

An activity is a sequence of operations to be performed by a certain type of resources. The Petri net model for the $k-th$ activity of resource r is described by a Petri net H^k_r that starts and ends with the resource idle state place p_r as follows.

Definition 3.4: Petri net $H^k_r(m^k_r)$ = (P^k_r , T^k_r , F^k_r , C^k_r , m^k_r) denotes the $k-th$ activity of resource r . There is no common transition between H_r^k and $H_r^{k'}$ for $k \neq k'$. The initial marking $m_r^k(p_r) = 1$ and $m_r^k(p) = 0$ for each $P_r^k \setminus \{p_r\}$, where p_r is the idle state place of resource r . $C_r^k(t) \cap C_r^k(t') = \Phi \forall t, t' \in T_r^k$ with $t \neq t'$.

Fig. 3 illustrates the resource activities associated with the workflows in Fig. 2. To solve TCSP based on the proposed TPN models we study the temporal property in the next section.

4 Temporal Property of Medical Workflows

To compute the earliest completion time for a workflow agent to execute transitions for given available time intervals from the resource agents, we study the temporal property of healthcare workflows by exploiting the workflow structure. We define a token flow path as follows.

Definition 4.1: A token flow path $\pi = p_1 t_1 p_2 t_2 p_3 t_3 .....p_n t_n p$ is a directed path of w_n . $\prod_n$ denotes the set of all token flow paths starting with the service input place and ending with the service output place in w_n .

Property 4.1: Given w_n under marking m_n , the shortest time for a token to arrive at the sink place θ_n from the source place ε_n is $\underline{\tau}_n = \max_{\pi \in \prod_n} f_\alpha(\pi)$, where $f_\alpha(\pi) = \sum_{t_n \in \pi} \alpha_n$. The longest time for a token to arrive at the sink place θ_n from the source place ε_n is $\overline{\tau}_n = \max_{\pi \in \prod_n} f_\beta(\pi)$, where $f_\beta(\pi) = \sum_{t_n \in \pi} \beta_n$.

Example 1: For w_4 in Fig. 2, we have Π ={ π_1 , π_2 }, where $\pi_1 = p_5 t_5 p_6 t_6 p_7 t_7 p_8 t_{10} p_{12}$, $\pi_2 = p_5 t_5 p_9 t_8 p_{10} t_9 p_{11} t_{10} p_{12}$. $\underline{\tau} = \max_{\pi \in \Pi} f_\alpha(\pi)$ = $\max(\alpha_5 + \alpha_6 + \alpha_7 + \alpha_{10}, \quad \alpha_5 + \alpha_8 + \alpha_9 + \alpha_{10}\}$. $\overline{\tau} = \max_{\pi \in \Pi} f_\beta(\pi)$ = $\max(\beta_5 + \beta_6 + \beta_7 + \beta_{10}, \beta_5 + \beta_8 + \beta_9 + \beta_{10}$.

Property 4.1 facilitates temporal reasoning efficiently based on TPN without applying reachability analysis method. Our problem method is based on Property 4.1. Let U_n denote the set of managers of workflow agent w_n . Let $i \in U_n$ and $\underline{\sigma}_i$ denote the earliest time for a token from workflow agent w_i to arrive at place ε_n , where w_i is a manager of w_n . Workflow agent w_n computes the earliest time and the latest time for a token to arrive at the sink place θ_n as follows.

Property 4.2: The earliest time for a token to arrive at place θ_n of a workflow agent w_n is $\underline{\sigma}_n = \min_{i \in U_n} \underline{\sigma}_i + \max_{\pi \in \prod_n} f_\alpha(\pi)$, where $f_\alpha(\pi) = \sum_{t_n \in \pi} \alpha_n$. The latest time for a token to arrive at the sink place θ_n from the source place ε_n is $\overline{\sigma}_n = \max_{i \in U_n} \overline{\sigma}_i + \max_{\pi \in \prod_n} f_\beta(\pi)$, where $f_\beta(\pi) = \sum_{t_n \in \pi} \beta_n$.

5 Agent Interaction Mechanism

Our approach to TCSP relies on interaction between patient agents, workflow agents and resource agents (including doctor agents, staff agents, specialist agents and nurse agents). Interactions among different types of agents are based on a negotiation mechanism that extends the well-known contract net protocol (CNP) [14]. CNP relies on an infrastructure for individual agents to publish and discover their services and communicate with each other based on the ACL language defined by the FIPA international standard for agent interoperability. To make a workflow agent discover other related agents, each workflow agent publishes the service input places and service output places. To illustrate our service discovery mechanism, consider the scenario of Fig. 2. For Fig. 2, workflow agents w_1 ~ w_9 publish their service input places via DF agent, which provides yellow page services to workflow agents. For Fig. 2, w_1, w_2, w_3, w_4, w_5, w_6, w_7, w_7 and w_9 publishes $\varepsilon_1 = p_1$, $\varepsilon_2 = p_3$, $\varepsilon_3 = p_5$, $\varepsilon_4 = p_5$, $\varepsilon_5 = p_{12}$, $\varepsilon_6 = p_{14}$, $\varepsilon_7 = p_{16}$, $\varepsilon_8 = p_{18}$ and $\varepsilon_9 = p_{20}$, respectively. To discover other agents' services provided, a workflow agent must search the yellow page services by providing the DF with a template description that specifies its service output place. The result of the search is the list of all the descriptions that match the provided template.

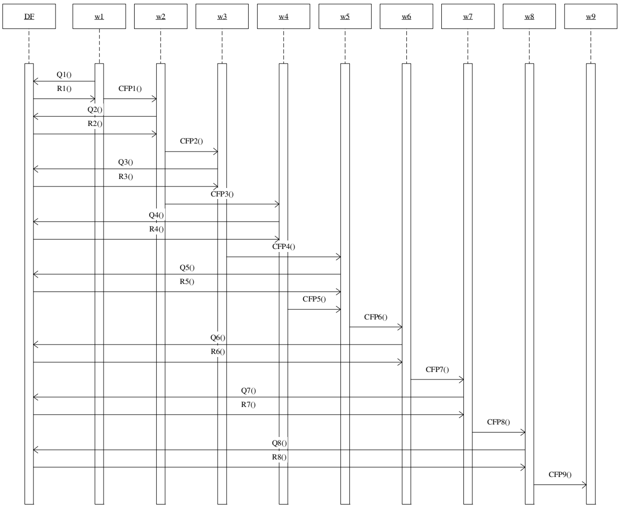


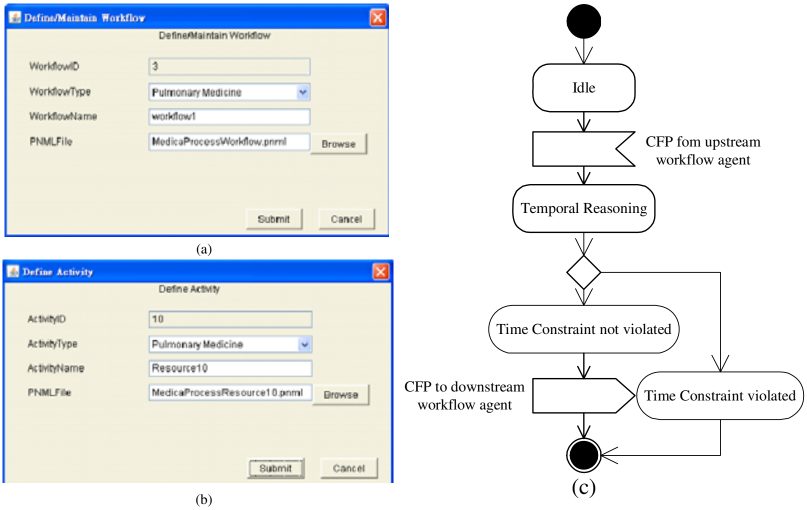
Fig. 4. Sequence diagram of service discovery and call for proposals

The problem solving process is initiated with the reception of a request to a workflow agent from the patient agent. Following reception of the request, the workflow agent negotiates with other agents (including workflow agents and resource agents) to attempt to create a solution. A workflow agent w_m with a service output place θ_m will query the DF Agent to discover any other workflow agent w_n with $\varepsilon_n = \theta_m$ and issue "Call for proposals" message (CFP) to them as needed. A workflow agent w_m will also query the DF Agent to discover resource agents required to perform the activities in workflow agent w_m. Fig. 4 shows how workflow agent w_1 searches the yellow page for the workflow agent with service input place p_3 by sending Q1. In this example, w_1 queries the DF Agent to discover w_2 and w_3 that can provide the requested services. Once w_3 and w_2 are discovered, w_1 will send a "Call for proposals" message (CFP1) to w_2 and a "Call for proposals" message (CFP2) to w_3 as shown in Fig.4. On receiving the "Call for proposals" message (CFP1), w_2 searches the yellow page by sending message Q2 as shown in Fig. 4 to discover the potential workflow agents with service input place p_5. A similar process follows. Fig. 4 illustrates the sequence diagram of service discovery and call for proposals. Based on the services publication and discovery mechanism, CNP can be applied to facilitate negotiation and temporal reasoning of workflow agents. Fig. 5 shows the screen shot to define a workflow and a activity and the state diagram to handle a CFP by a workflow agent. The following polynomial complexity algorithm is applied by workflow agent w_n to find the earliest completion $\underline{\sigma}_n$ and the latest completion time $\overline{\sigma}_n$ based on the earliest completion time $\underline{\sigma}$ and the latest completion time $\overline{\sigma}$ of the manager.

```
Temporal Reasoning Algorithm for workflow agent $w_n$
Input:  $\underline{\sigma}$ : The earliest completion time of the manager
        $\overline{\sigma}$ : The latest completion time of the manager
Output: $\underline{\tau}_n$ : The shortest processing time of agent $w_n$
        $\overline{\tau}_n$ : The longest processing time of agent $w_n$
        $\underline{\sigma}_n$ : The earliest completion time of agent $w_n$
        $\overline{\sigma}_n$ : The latest completion time of agent $w_n$
Step 1: Find the shortest processing time by $\underline{\tau}_n = \max_{\pi\in\prod_n} f_\alpha(\pi)$, where $f_\alpha(\pi) = \sum_{t_n\in\pi} \alpha_n$
Step 2: Find the longest processing time by $\overline{\tau}_n = \max_{\pi\in\prod_n} f_\beta(\pi)$, where $f_\beta(\pi) = \sum_{t_n\in\pi} \beta_n$.
Step 3: Find the earliest completion time of agent $w_n$ by $\underline{\sigma}_n = \underline{\sigma} + \underline{\tau}_n$
        Find latest completion time of agent $w_n$ by $\overline{\sigma}_n = \overline{\sigma} + \overline{\tau}_n$
```

For Fig. 2, the earliest completion time is $\alpha_1 + \alpha_2 + \alpha_3 + \alpha_4 + \min(\max(\alpha_5 + \alpha_6 + \alpha_7 + \alpha_{10},\ \alpha_5 + \alpha_8 + \alpha_9 + \alpha_{10}), \alpha_{21} + \alpha_{22}) + \alpha_{11} + \alpha_{12} + \alpha_{13} + \alpha_{14} + \alpha_{15} + \alpha_{16} + \alpha_{17} + \alpha_{18} + \alpha_{19} + \alpha_{20}$. The latest completion time is $\max(\beta_1 + \beta_2 + \beta_3 + \beta_4 + \beta_5 + \beta_6 + \beta_7 + \beta_{10} + \beta_{11} + \beta_{12} +$

$$\beta_{13}+\beta_{14}+\beta_{15}+\beta_{16}+\beta_{17}+\beta_{18}+\beta_{19}+\beta_{20},$$
$$\beta_{1}+\beta_{2}+\beta_{3}+\beta_{4}+\beta_{5}+\beta_{8}+\beta_{9}+\beta_{10}+\beta_{11}+\beta_{12}+$$
$$\beta_{13}+\beta_{14}+\beta_{15}+\beta_{16}+\beta_{17}+\beta_{18}+\beta_{19}+\beta_{20},$$
$$\beta_{1}+\beta_{2}+\beta_{3}+\beta_{4}+\beta_{21}+\beta_{22}+\beta_{13}+\beta_{14}+\beta_{15}+\beta_{16}+\beta_{17}+\beta_{18}+\beta_{19}+\beta_{20}.$$



Define/Maintain Workflow

WorkflowID	3
WorkflowType	Pulmonary Medicine
WorkflowName	workflow1
PNMLFile	MedicaProcessWorkflow.pnml

(a)

Define Activity

ActivityID	10
ActivityType	Pulmonary Medicine
ActivityName	Resource10
PNMLFile	MedicaProcessResource10.pnml

(b)

Idle
CFP fom upstream workflow agent
Temporal Reasoning
Time Constraint not violated
CFP to downstream workflow agent
Time Constraint violated

(c)

Fig. 5. (a)Define workflow (b)Define activity (c)State diagram to handle a CFP

6 Conclusion

The problem to determine whether the medical treatment of a patient can be completed by a time constraint using the available resources in a hospital is formulate as a Temporal Constraint Satisfaction Problem (TCSP). We propose a viable and systematic approach to develop a problem solver for TCSP in hospitals based on MAS. We study how to estimate the shortest completion and the longest completion time in the presence of uncertainty in operation time and provide a seamlessly framework to guide the healthcare workers to take the next actions based on the medical processes. Our approach combines MAS with formal TPN models to solve TCSP. We propose a subclass of TPN models to describe the workflows of workflow agents and the activities of resource agents in hospitals. Uncertainty in timing is represented by static interval of a transition in TPN. Existing analysis methods for TPN models such as the one proposed in [16, 23] based on state class graph (SCG) cannot be applied directly to a distributed MAS environment. In this paper, we proposed a temporal analysis method based on MAS architecture to efficiently deduce its earliest and latest completion time based on the timing information from other agents. Our temporal reasoning algorithm is developed based on MAS architecture. Construction of SCG is not required as our analysis method is developed by exploiting the structure of the TPN. The computational complexity of our analysis algorithm is polynomial. Therefore, our algorithm is much more efficient than SCG for the subclass of TPN. It is also scalable for large scale problems. Our results indicate that reasoning about temporal constraints in the subclass of TPN proposed in this paper can be achieved in polynomial time by exploiting its structure.

Acknowledgment. This paper is supported in part by National Science Council of Taiwan under Grant NSC100-2410-H-324-006-MY2.

References

1. Decker, K., Li, J.: Coordinated hospital patient scheduling. In: International Conference on Multi Agent Systems, Paris, July 3-7, pp. 104–111 (1998)
2. Kutanoglu, E., David Wu, S.: On combinatorial auction and Lagrangean relaxation for distributed resource scheduling. IIE Transactions 31, 813–826 (1999)
3. Oddi, A., Cesta, A.: Toward interactive scheduling systems for managing medical resources. Artificial Intelligence in Medicine 20, 113–138 (2000)
4. Daknou, A., Zgaya, H., Hammadi, S., Hubert, H.: A dynamic patient scheduling at the emergency department in hospitals. In: IEEE Workshop on Health Care Management, Venice, February 18-20, pp. 1–6 (2010)
5. Spyropoulos, C.D.: AI planning and scheduling in the medical hospital environment. Artificial Intelligence in Medicine 20, 101–111 (2000)
6. Nilsson, N.J.: Artificial Intelligence: A New Synthesis. Morgan Kaufmann Publishers, Inc., SanFrancisco (1998)
7. Ferber, J.: Multi-Agent Systems, An Introduction to Distributed Artificial Intelligence. Addison-Wesley, Reading (1999)
8. Durfee, E.H., Lesser, V.R., Corkill, D.D.: Trends in cooperative distributed problem solving. IEEE Transactions on Knowledge and Data Engineering 1(1), 63–83 (1989)
9. Russel, S.J., Norvig, P.: Artificial Intelligence—A Modern Approach, 2nd edn. Pearson Education Asia Limited (2006)
10. Workflow Management Coalition, XPDL support and resources (2009), `http://www.wfmc.org/xpdl.html`
11. Object Management Group, Business process modeling notation (2009), `http://www.bpmn.org`
12. OASIS, Web services business process execution language version 2.0 (2009), `http://docs.oasis-open.org/wsbpel/2.0/OS/wsbpel-v2.0-OS.html`
13. Merlin, P., Farbor, D.: Recoverability of communication protocols. IEEE Trans. on Communications 24(9), 1036–1043 (1976)
14. Smith, R.G.: The Contract net protocol: high-level communication and control in a distributed problem solver. IEEE Transactions on Computers 29, 1104–1113 (1980)
15. Murata, T.: Petri Nets: Properties, Analysis and Applications. Proceedings of the IEEE 77(4), 541–580 (1989)
16. Berthomieu, B., Menasche, M.: An Enumerative Approach for Analyzing Time Petri Nets. In: Proc. Ninth International Federation of Information Processing (IFIP) World Computer Congress, vol. 9, pp. 41–46 (September 1983)
17. Conry, S.E., Kuwabara, K., Lesser, V.R., Meyer, R.A.: Multistage negotiation for distributed constraint satisfaction. IEEE Transactions on Systems, Man and Cybernetics 21(6), 1462–1477 (1991)

Assessing Rice Area Infested by Brown Plant Hopper Using Agent-Based and Dynamically Upscaling Approach

Vinh Gia Nhi Nguyen[1], Hiep Xuan Huynh[2], and Alexis Drogoul[1]

[1] UMI 209 UMMISCO, IRD/UPMC, Institut de recherche pour le développement
Bondy, France
{nngvinh,alexis.drogoul}@gmail.com
[2] DREAM Team, School of Information and Communication Technology,
Cantho Univesity, Vietnam
hxhiep@ctu.edu.vn

Abstract. This paper introduces an agent-based framework to modeling and simulate assessing rice area infested by brown plant hoppers from field to regional scales using agent-based and upscaling approach. Because detail levels of infestation rice area information are different from small scale to large scale, infestation rice area information are collected from fields in the Mekong Delta region-Vietnam and are upscaled to province scale, results are validated by province estimation reports. From the results, infestation rice area information can be used as an indicator for assessing damage levels of rice crop seasons caused by Brown plant hoppers, agent-based model and uscaling method could be an usefull tool for aggregating information from field to decision-makers and support planning argicultural strategics.

Keywords: Brown plant hopper (BPH), Scaling, Upscaling, Agent-based model, GAMA platform.

1 Introduction

Brown plant hopper is one kind of Nilaparvata lugens species and is the major pest on rice [2],[5], this insect damaged rice crop systems in Asia and carries rice virus-related diseases like grassy stunt, ragged stunt and wilted stunt, etc, that causing low rice product. The local and national government tried to control BPHs invasion and outbreaks in the Mekong Delta region-Vietnam through activities of researchers and farmers who collected BPH data by making observation [8]. Questions raised as exploring Brown plant hopper problems: “How to predict invasion of brown plant hoppers?” or “How to know BPH density at certain location in time and space in the Mekong Delta?”, controlling flight activity and population dynamic of Brown plant hoppers was conducted by directly sampling surveys and by trapping BPH adults during rice seasons. Reseachers and staffs from plant protection department at commune level collect daily BPH density (individuals/m^2) and BPH infestation rice area (ha) and send weekly reports to plant protection department at district level, a estimation synthesis of reports is made by staffs and distribute to decision-makers at province scale to have suitable policy during rice crop season [10]. At field scale, almost farmers based on individual experience and

J.-S. Pan, S.-M. Chen, N.T. Nguyen (Eds.): ACIIDS 2012, Part I, LNAI 7196, pp. 43–52, 2012.

recommendation messages from local government at district and province level to choose which kind of rice crops and fertilisation. However, BPH infestation rice area data are available at field scale and are more detail but not available at district and province scale, BPH infestation rice area information needed in upper scale are more general in meaning. Fig.1 illustrates information flows between administrative scales:

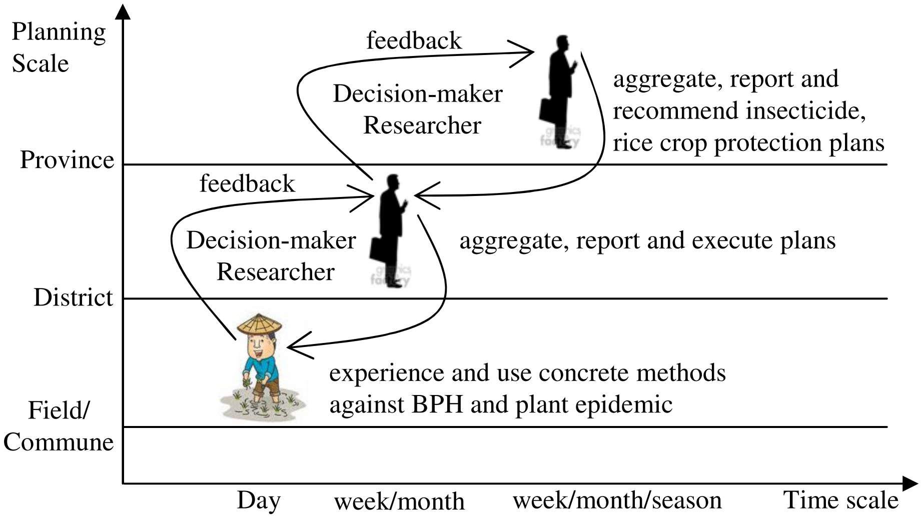

Planning Scale
feedback
Decision-maker
Researcher
aggregate, report and recommend insecticide, rice crop protection plans
Province
feedback
Decision-maker
Researcher
aggregate, report and execute plans
District
experience and use concrete methods against BPH and plant epidemic
Field/ Commune
Day
week/month
week/month/season
Time scale

Fig. 1. Scale schema at space and time

Scaling is the transmission of information between scales [1] and scaling can help filling gap between one scale at which most information is willing and other scale at which most decisions concerning plans are made [3]. Aim of upscaling is to answer questions at scales that can not be solved directly, e.g. decision-makers at province scale no need micro information at field/commune scale but they need macro information from district scale to make long term provincial strategies and prediction in at least one week/month/season/year. Almost macro information at district and province scale are estimated by experience and knowledge of experts. Thus, an agent-based model [4],[6] is necessary to capture real world objects: decision-maker, researcher, farmer, BPH, etc as agents and to make linkages through reports, experiences and feedbacks between agents for assessing BPH infestation rice area. This paper attempts to describe a model using agent-based and upscaling approach to assess BPH infestation rice area in the hierarchy in Fig.1 from field/commune scale to province scale. The paper is structured as follows: section 2 proposes MSMEKO model using agent-based and upscaling approach, section 3 presents results and discussion and section 4 draws conclusion.

2 MSMEKO Model Using Agent-Based and Upscaling Approach

MSMEKO model is built on GAMA platform [7]. GAMA supports GIS tools to design spatial environment for agents, status of agents can be explored during simulation. Active agents such as decision-maker, BPH, rice have attributes and actions (Fig.2), other passive agents could be weather and spatial objects from GIS data (province, district, commune). Relationship lines in Fig.2 determine possible inheritance and interactions among agents. A land use GIS map data is loaded into simulation environment, agents like province, commune and district will be created automatically with attributes: code, name, area and coordinates. Because of time and large memory consuming while simulating for each BPH individual in land use GIS map, a group of BPHs could be considered as an agent and automatically updates number of eggs, nymphs, adults and number of invading adults to other group agents at each simulation step. Real world objects in the Mekong Delta region are represented as agents in computer model in the following table:

Table 1. Decision-maker, insect, vegetation and environmental components

Agent name	Description
Decision-maker and Organization	
Farmer	farmer make rice crops in the fields and represents household unit level
PdecisionMaker, DdecisionMaker, CdecisionMaker	officers are in charge of choosing and planning rice crop systems and other protection policies that help and improve crop production, e.g. Plant protection department in each level: province (P), district (D), commune (C)
Environment	
Weather	weather variables and factors affecting BPH cycle life: rainfall, temperature, humidity, wind, ….
LightTrap	a tool used to trap BPH adults, each commune has at least one light trap
LandUnit	smallest land unit or a field to make rice crop
LandFarm	a collection of LandUnits is owned by a farmer
Commune	a collection of LandFarms is belong to commune
District	a set of communes in a district
Province	a set of districts in a province
Insect	
BPH	Brown plant hopper individual has biological characteristics and perceptions with environment. Growth of BPH depends on rice growth stages
GBph	Group of Brown plant hopper in a rice field
Vegetation	
Rice	main crop is planted in the Mekong delta region

Attributes and processes in each agent will be adapted during implementing phase. *CdecisionMaker* agent uses action *report()* to collect daily BPH density and BPH infestation rice area from sample rice fields and send weekly reports to *DdecisionMaker* agent at district scale, these reports are synthesized by action *aggregate()* and action *report()* to sent weekly to *PdecisionMaker* agent for decisions concerning agricultural systems. This paper focused on aggregating/upscaling information function for conveying BPH infestation rice area infromation from *CdecisionMaker* agent to *PdecisionMaker* agent.

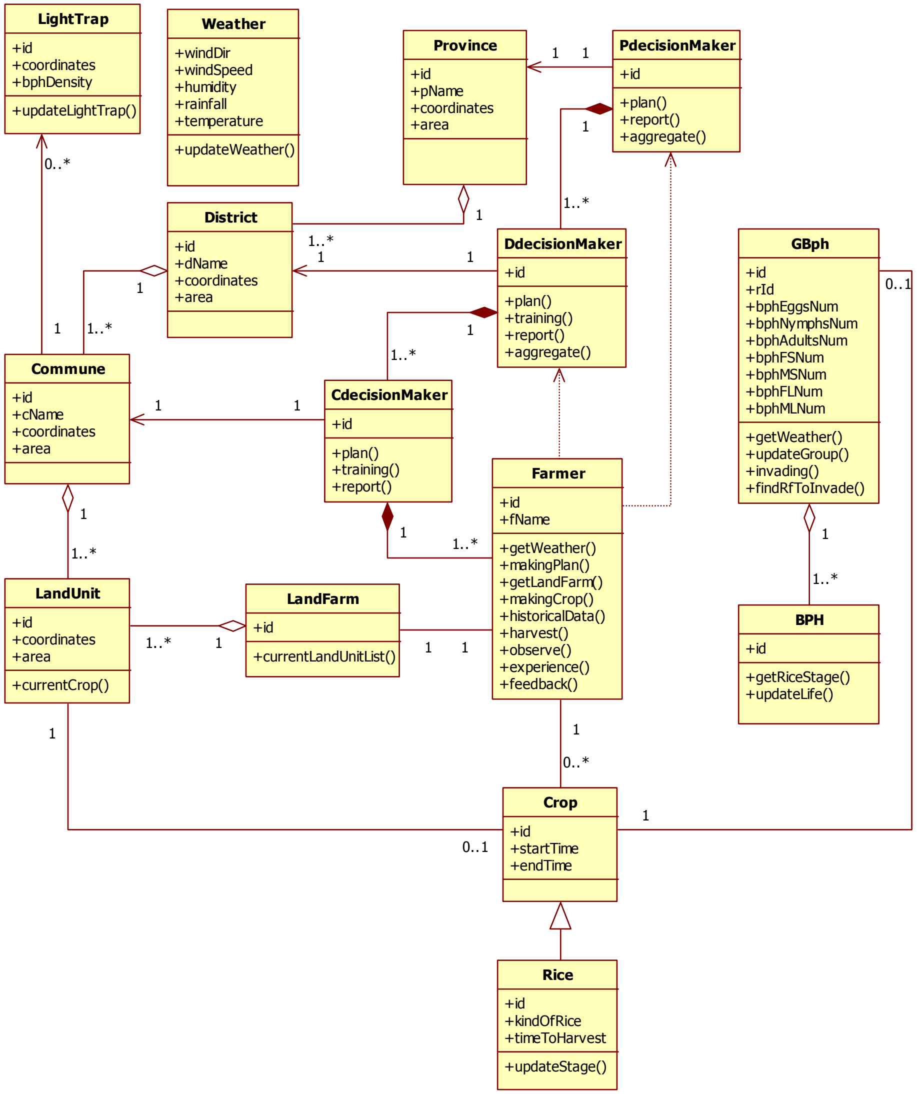


Fig. 2. UML static diagram of objects abstraction

Table 2. BPH infestation levels

BPH infestation level	Legend	Description (individuals/m²)
Level 1	Light BPH density	BPH density >0 and <1500
Level 2	Medium BPH density	BPH density >=1500 and <3000
Level 3	Heavy BPH density	BPH density >=3000

To assess status of BPH infestation at commune scale, BPH infestation in the study [9] are classified into three levels: light, medium and heavy infestation (table 2). An example of BPH infestation levels at commune scale on day 01-July-2009 in *Dong Thap* province below:

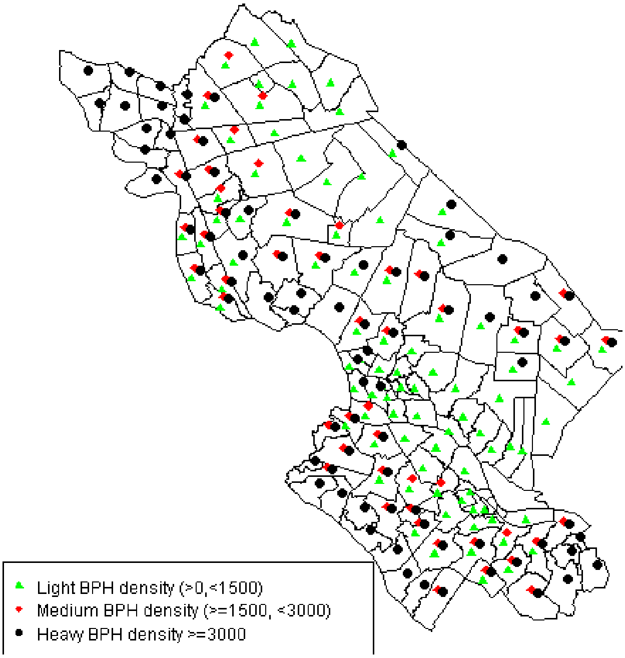

▲ Light BPH density (>0,<1500)
◆ Medium BPH density (>=1500, <3000)
● Heavy BPH density >=3000

Fig. 3. Three BPH infestation levels on rice fields in *Dong Thap* province, on 01-July-2009

Due to National technical regulation on surveillance methods of plant pests [11], model at field/commune scale and day time scale (Fig.4) is aggregating function of daily BPH density data at survey locations. Eq.1 computes BPH density (individuals/m^2) information ad^i in each commune at ith level:

$$ad^i = \frac{1}{S^i}\sum_{k=1}^{h^i} x_k \qquad (1)$$

where: x_k is number of BPHs at kth survey location, S^i is total survey rice area (m^2) having ith level (table 2), h^i is number of survey locations having ith level.

BPH infestation rice area at ith level in each commune are modeled by:

$$c^i(ha) = \frac{n^i}{n} S \qquad (2)$$

with n_i is number of survey locations having ith level, n is number of survey locations, S is total rice area of a commune, c^i is BPH infestation rice area at ith level.

In upscaling approach (Fig.4), at district scale, BPH infestation rice area information is aggregated from BPH infestation rice area information that is collected from field/ commune scale within real land use GIS maps of regional scale. BPH infestation rice area of a district d^i at ith level is aggregated from BPH infestation rice areas of **m** communes:

$$d^i = \mathrm{w}_1 * \mathrm{c}_1^{\mathrm{i}} + \mathrm{w}_2 * \mathrm{c}_2^{\mathrm{i}} + \ldots + \mathrm{w}_{\mathrm{m}} * \mathrm{c}_{\mathrm{m}}^{\mathrm{i}} = \sum_{j=1}^{m} w_j * \mathrm{c}_{\mathrm{j}}^{\mathrm{i}} \qquad (3)$$

This is also similar when aggregating BPH infestation rice area information from district to province level.

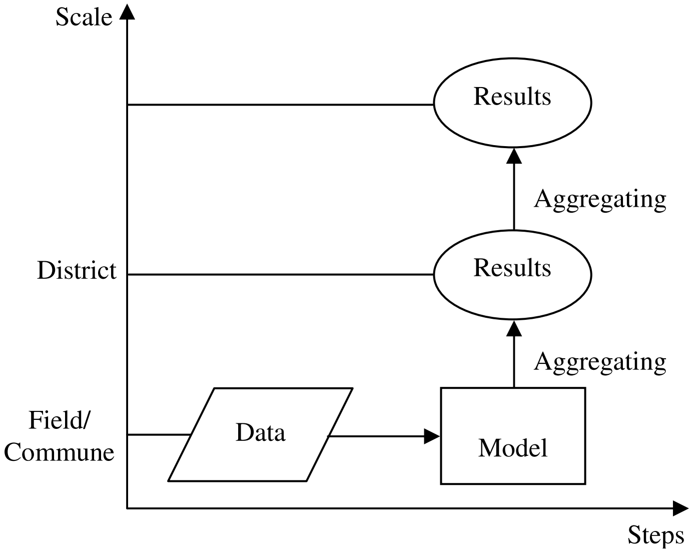


Fig. 4. Upscaling BPH infestation rice area information from field to province scale

An assumption in this paper is that there is relationships or influences between fields/communes and values w_1, w_2, w_3, ..., w_m can be considered as weights or influence factors, and values $\mathrm{c}_1^{\mathrm{i}}$, $\mathrm{c}_2^{\mathrm{i}}$, ..., $\mathrm{c}_\mathrm{m}^{\mathrm{i}}$ are BPH infestation rice area obtained from **m** communes (Eq.2). Weighting function **w** is used to determine weight values with arameters that distinguish between communes, function **w** could be implemented in different ways according to problems. Choosing aggregating function $^{\mathbf{i}}$ and function **w** requires discussion between decision-maker, researcher and farmer because results from upscaling procedure could be used as main reference for policy planning. Three kinds of popular questions about brown plant hopper problems are mostly needed by decision-makers at district and province scales as follows:

(Q1). what is total BPH infestation rice area for each level (table 2) at district/province scale?

(Q2). what is average BPH infestation rice area for each level (table 2) per district at province scale?

(Q3). what is total BPH infestation rice area for each level (table 2) on summer-autumn rice crops in July at district/province scale?

To solve questions, in MSMEKO model, three upscaling procedures are proposed to understand influence of weight values to process of aggregating BPH infestation rice area information, each procedure tries to answer a question:

Procedure 1. (question **Q1**): communes have the same influence each other, weight values are equal to value 1:

$$w_j = 1, j = 1,2,...,m \quad \text{==>} \quad d^i = \sum_{j=1}^{m} c_j^i \tag{4}$$

m: number of communes in a district

Procedure 2. (question **Q2**): weight values of each commune are equal to $\frac{1}{m}$, BPH infestation rice area information at district scale is:

$$w_j = \frac{1}{m} \quad , j = 1,2,...,m \quad \text{==>} \quad d^i = \frac{1}{m}\sum_{j=1}^{m} c_j^i \tag{5}$$

Procedure 3. (question **Q3**): because July not only is ending time of summer-autumn (*SA*) rice season but also is starting time of autumn-winter (*AW)* rice season, thus let $c_j^i = c_{j,SA}^i + c_{j,AW}^i$ is sum of BPH infestation rice area on *SA* rice crops and BPH infestation rice area on *AW* rice crops at i^th^ infestation level, weight values are: $w_j = w_{j,SA} + w_{j,AW}$. Total BPH infestation rice area at i^{th} level in district d^i is:

$$d^i = \sum_{j=1}^{m} \left(w_{j,SA} * c_{j,SA}^i + w_{j,AW} * c_{j,AW}^i \right)$$

Function $\mathbf{d^i}$ in Eq.6 is non-linear function to solve questions like (**Q3**):

$$\left[\begin{array}{ll} d^i = \sum_{j=1}^{m} w_{j,SA} * c_{j,SA}^i & \text{if rice season in question (}\mathbf{Q3}\text{) is summer-autumn} \\ d^i = \sum_{j=1}^{m} w_{j,AW} * c_{j,AW}^i & \text{if rice season is autumn-winter.} \end{array}\right.$$

The model can be extended to region scale from province scale. For example, what is total BPH infestation rice area in a month/season in a certain region?. Agents *PdecisionMaker*, *DdecisionMaker* and *CdecisionMaker* could assess BPH infestation rice area information at each scale by using above aggregating functions as actions.

3 Results and Discussion

The MSMEKO model is experimented on land use GIS map, BPH density and BPH infestation rice area data in *Dong Thap* province in the Mekong Delta region, South of Vietnam. It was located in 10°07'14"-10°58'18" north latitude and 105°11'38"-105°56'42" east longitude, at the average elevation of approximately 2 meters above sea level. There were 134 communes in 11 districts making rice crops during a year. The total area is 3246 km^2 including 408.294 ha for farming rice crops and 245.044 ha for doing other crops. Data used for simulation are daily BPH density and weekly BPH infestation rice area in communes on autumn-winter rice season in 2009.

The results of aggregating BPH infestation rice area from commune to district scale are described in Fig.5 for upscaling procedure 1 (Eq.4), each circle with a color reflects BPH infestation rice area at different infestation levels (table 2). It is possibility for geographical mapping analysis (Fig.5) to help decision-makers detect BPH hot-spots (large BPH infestation rice area) in communes or districts, e.g. heavy BPH infestation rice area in *Sa Dec* and *Tan Hong* district is very high in comparision with other districts. In the 1st week (Fig.5), BPH infestation rice area in *Lap Vo*, *Lai Vung* and *Chau Thanh* district have the same circle perimeter of light and medium BPH infestation rice area although they have different rice crop areas, this helps decision-makers find out district groups or clusters having similar BPH infestation rice area.

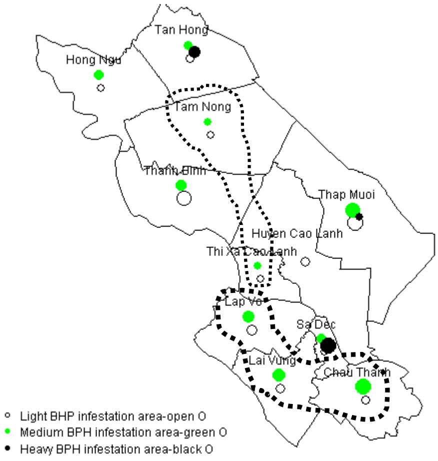

Tan Hong
Hong Ngu
Tam Nong
Thanh Binh
Thap Muoi
Huyen Cao Lanh
Thi Xa Cao Lanh
Lap Vo
Sa Dec
Lai Vung
Chau Thanh
o Light BHP infestation area-open O
• Medium BPH infestation area-green O
• Heavy BPH infestation area-black O

Fig. 5. District BPH infestation rice area is upscaled from communes in Dong Thap province on the 1st week, July, 2009 using procedure 1

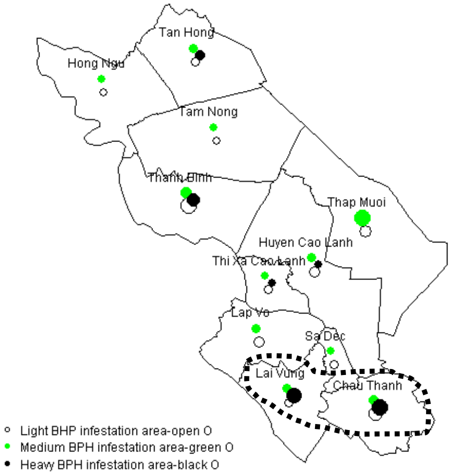

Tan Hong
Hong Ngu
Tam Nong
Thanh Binh
Thap Muoi
Huyen Cao Lanh
Thi Xa Cao Lanh
Lap Vo
Sa Dec
Lai Vung
Chau Thanh
o Light BHP infestation area-open O
• Medium BPH infestation area-green O
• Heavy BPH infestation area-black O

Fig. 6. District BPH infestation rice area is upscaled from communes in Dong Thap province on the 2sd week, July, 2009 using procedure 1

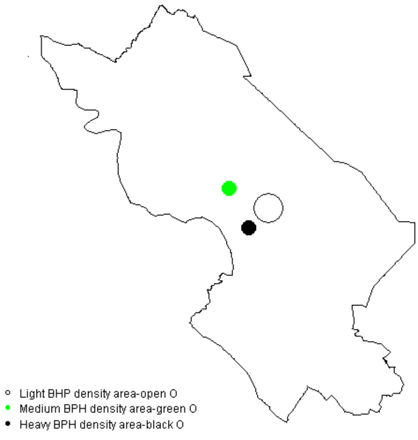

o Light BHP density area-open O
• Medium BPH density area-green O
• Heavy BPH density area-black O

Fig. 7. Province BPH infestation rice area is upscaled from districts on the 1st week – July, 2009 with procedure 1.

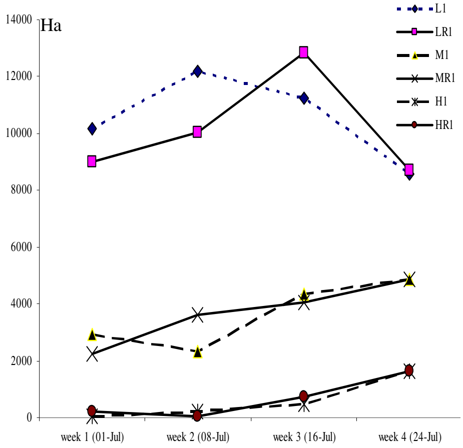


Fig. 8. Total BPH infestation rice area for each level at province scale, weekly, July-2009. L1, M1, H1 are total BPH infestation rice area of light, medium and heavy levels at province scale.

In Fig.5, *Huyen Cao Lanh* district has light infestation rice area, after one week (Fig.6) this district occurs small medium and heavy infestation rice area.

In the 2sd week of July (Fig.6), because almost districts finished harvesting summer-autumn rice crops and started autumn-winter rice season, *Sa Dec* district has low remaining rice area in ripening stage and BPHs migrate to other rice fields. Similarly, Fig.7 shows BPH infestation rice area at province scale is aggregated from BPH infestation rice areas at district scale. Total BPH infestation rice area (ha) at province scale using procedure 1 (Fig.8) denote that results from simulation and from estimation reports have some differences, and LR1, MR1, HR1 are estimated results of total light, medium and heavy infestation (ha) from estimation reports [10].

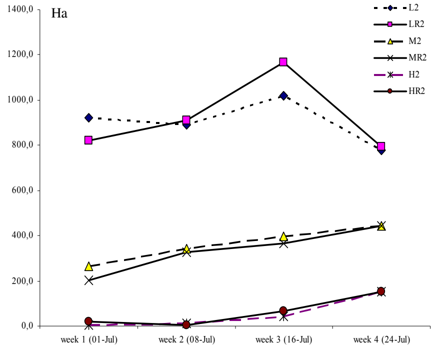


Fig. 9. Average BPH infestation rice area for each level at province scale, weekly, July-2009. L2, M2, H2 are average BPH infestation rice area of light, medium and heavy levels per district at province scale

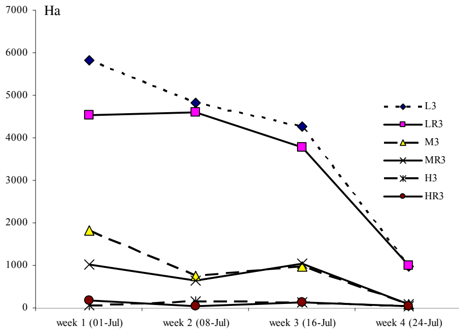


Fig. 10. Total BPH infestation rice area for each level on summer-autumn rice crops at province scale, weekly, July-2009. L3, M3, H3 are total BPH infestation rice area of light, medium and heavy levels at province scale

Fig.9 shows weekly average BPH infestation rice area per district resulted from upscaling procedure 2 (Eq.5), results is rather fitted with infestation rice area indicator in estimation reports (LR2, MR2, HR2). In Fig.10, procedure 3 (Eq.6) shows that at the end of July summer-autumn rice season is mostly finished, BPH infestation rice area for this season descreased as rapidly as that in estimated reports at province level (LR3, MR3, HR3). An extension of upscaling from province scale could be applied to Mekong delta region scale that includes province agents in assessing BPH infestation rice area cover all province agents. Regional plant protection center bases on estimation reports from provinces to have a general understanding on BPH infestation rice area of provinces and have more suitable policies between provinces.

4 Conclusion

MSMEKO model tried to use upscaling approach as a tool to simulate conveying BPH infestation rice area information from field to region scale and help understand macro assessment information at region scale. Visualizing geographical maps like Fig.6 and Fig.7 is much difficult because of lacking data in reality at upper scale, thus upscaling methods are necessary to help understanding scale at which information are not available. Choosing an appropriate aggregating function at each scale requires an circle discussion process between three objects: decision-maker, farmer and researcher. Along with experiences, decision-makers will make prediction, protection plans and send recommendation reports to farmers for improving rice crops.

Acknowledgments. Authors wish to thank staffs in Land resources department, School of Environment and Natural Resources, Cantho university for supporting land use GIS data. We appreciate staffs of Dong Thap plant protection department who supported and provided BPH density and BPH infestation rice area data in Dong Thap province. Finally, this publication was made possible through support provided by the IRD-DSF.

References

1. Bierkens, F.P., Finke, P.A., de Willigen, P.: Upscaling and downscaling methods for environmental research. Kluwer Academic Publishers (2000)
2. Cheng, J.: Rice planthopper problems and relevant causes in China. In: Heong, K.L., Hardy, B. (eds.) Planthoppers: New Threats to the Sustainability of Intensive Rice Production Systems in Asia, pp. 157–178. The International Rice Research Institute (IRRI), Los Baños (2009)
3. Dalgaard, T., Hutchings, N.J., Porter, J.R.: Agroecology, scaling and interdisciplinarity. Agriculture, Ecosystems & Environment 100, 39–51 (2003)
4. Damgaard, M., Kjeldsen, C., Sahrbacher, A., Happe, K., Dalgaard, T.: Validation of an agent-based, spatio-temporal model for farming in the RIver Gudenå landscape. Results from the MEA-Scope case study in Denmark. In: Piorr, A., Müller, K. (eds.) Rural Landscapes and Agricultural Policies in Europe, pp. 241–258. Springer, Berlin (2009)
5. Dyck, V.A., Thomas, B.: The brown planthopper problem. In: Brown Planthopper: Threat to Rice Production in Asia, pp. 3–17. The International Rice Research Institute, Los Baños (Philippines) (1997)
6. Drogoul, A., Vanbergue, D., Meurisse, T.: Multi-agent Based Simulation: Where Are the Agents? In: Sichman, J.S., Bousquet, F., Davidsson, P. (eds.) MABS 2002. LNCS (LNAI), vol. 2581, pp. 1–15. Springer, Heidelberg (2003)
7. Drogoul, A., et al.: Gama simulation platform (2011), `http://code.google.com/p/gama-platform/`
8. Heong, K.L., Escalada, M.M., Huan, N.H., Chien, H.V., Choi, I.R., Chen, Y., Cabunagan, R.: Research and implementation issues related to management of the brown planthopper/virus problem in rice in Vietnam. In: Australian Centre for International Agricultural Research-ACIAR, vol. 08 (2008)
9. Phan, C.H., Huynh, H.X., Drogoul, A.: An agent-based approach to the simulation of brown plant hopper invasions (BPH) in the the Mekong Delta. In: IEEE-RIVF, pp. 227-232 (2010)
10. Provincial plant protection department (2009), `http://www.dongthap.gov.vn/wps/wcm/connect/Internet/sitbaodientu/sitathongtincanbiet/sitasinhvatgayhai/`
11. Plant protection department.: National technical regulation on Surveillance method of plant pests. Ministry of Agriculture and Rural Development-Vietnam, QCVN 01-38 (2010)

A Service Administration Design Pattern for Dynamically Configuring Communication Services in Autonomic Computing Systems

Vishnuvardhan Mannava[1] and T. Ramesh[2]

[1] Department of Computer Science and Engineering,
K.L. University, Vaddeswaram, 522502, A.P., India
vishnu@kluniversity.in
[2] Department of Computer Science and Engineering,
National Institute of Technology,
Warangal, 506004, A.P., India
rmesht@nitw.ac.in

Abstract. Rapidly growing collection of communication services is now available on the Internet. A communication service is a component in a server that provides capabilities to clients. Services available on the Internet include: WWW browsing and content retrieval services software distribution service. A common way to implement these services is to develop each one as a separate program and then compile, link, and execute each program in a separate process. However, this "static" approach to configuring services yields inflexible, often inefficient, applications and software architectures. The main problem with static configuration is that it tightly couples the implementation of a particular service with the configuration of the service with respect to other services in an application. In this paper we propose a system for dynamically configuring communication services. Server will invoke and manage services based on time stamp of service. The system will reduce work load of sever all services in executed by different threads based on time services are executed, suspended and resumed. Different patterns are used designing of service administration pattern that are reflective monitoring, strategy and thread per connection. This paper satisfies the properties of autonomic system: For monitoring use reflective monitoring, Decision making we use strategy pattern. Thread per connection is used of executing service in different thread. The pattern is described using a java-like notation for the classes and interfaces. A simple UML class and Sequence diagrams are depicted.

Keywords: Design Patterns, Dynamic Services and Autonomic System.

1 Introduction

The Service administration design pattern decouples the implementation services from the time at which the services are configured into an application or a system. This decoupling improves the modularity of services and allows the services to evolve over time independently of configuration issues (such as whether two services must

J.-S. Pan, S.-M. Chen, N.T. Nguyen (Eds.): ACIIDS 2012, Part I, LNAI 7196, pp. 53–63, 2012.

be co-located or what concurrency model will be used to execute the services) [1]. In addition, the Service administration pattern centralizes the administration of the services it configures. This facilitates automatic initialization and termination of services and can improve performance by factoring common service initialization and termination patterns into efficient reusable components.

Distributed computing applications grow in size and complexity in response to increasing computational needs, it is increasingly difficult to build a system that satisfies all requirements and design constraints that it will encounter during its lifetime. Many of these systems must operate continuously, disallowing periods of downtime while humans modify code and fine-tune the system [2]. For instance, several studies document the severe financial penalties incurred by companies when facing problems such as data loss and data inaccessibility. As a result, it is important for applications to be able to self-reconfigure in response to changing requirements and environmental conditions.

This approach enables a decision making process to dynamically evolve reconfiguration plans at run time. Autonomic systems sense the environment in which they are operating and take action to change their own behavior or the environment with a minimum effort. Every autonomic system having four properties that are monitoring, decision making, reconfiguration and execution, figure 1 will shows the autonomic system [5]

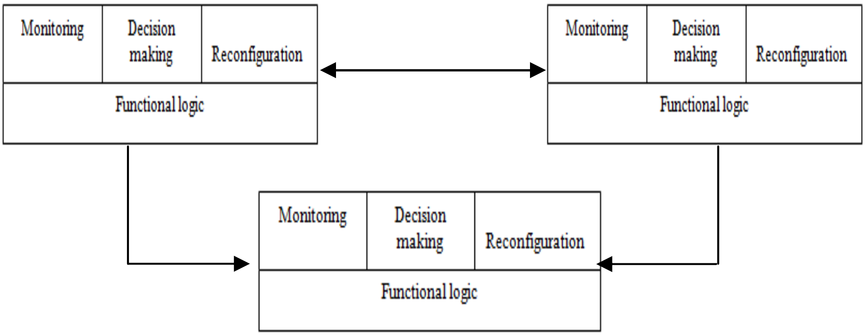


Fig. 1. Autonomic computing

2 Traps and Pitfalls with the Common Solution

Although the use of patterns like Reactor, Acceptor, and Connector improve the modularity and portability of the distributed time server, configuring communication services using a static approach has the following drawbacks [1]:

Service Configuration Decisions Must Be Made Too Early in the Development Cycle: This is undesirable since developers may not know *a priori* the best way to co-locate or distribute service components. For example, the lack of memory resources in wireless computing environments may force the split of Client and Clerk into two independent processes running on separate hosts. In contrast, in a real-time avionics environment it might be necessary to co-locate the Clerk and Server into one process to reduce communication latency. Forcing developers to commit prematurely to a particular service configuration impedes flexibility and can reduce performance and functionality [1].

Modifying a Service May Adversely Affect Other Services: The implementation of each service component is tightly coupled with its initial configuration. This makes it hard to modify one service without affecting other services. For example, in the real-time avionics environment Mentioned above, a Clerk and a Time Server might be statically configured to execute in one process to reduce latency. If the distributed time algorithm implemented by the Clerk is changed, however, the existing Clerk code would require modification, recompilation, and static relining. However, terminating the process to change the Clerk code would terminate the Time Server as well. This disruption in service may not be acceptable for highly available systems.

System Performance May Not Scale Up Efficiently: Associating a process with each service ties up OS resources (such as I/O descriptors, virtual memory, and process table slots). This design can be wasteful if services are frequently idle. Moreover, processes are often the wrong abstraction for many short-lived communication tasks (such as asking a Time Server for the current time or resolving a host address request in the Domain Name Service). In these cases, multithreaded Active Objects or single-threaded Reactive event loops may be more efficient.

3 Service Administration Design Pattern Template

To facilitate the organization, understanding, and application of the adaptation design patterns, this paper uses a template similar in style to that used by Ramirez et al. [2]. Likewise, the Implementation and Sample Code fields are too application-specific for the design patterns presented in this paper.

3.1 Pattern Name

Service administration design pattern

3.2 Classification

Structural - Monitoring

3.3 Intent

System deals with service invocation and management, when client invoke service from existing services observer will observe and report to service class it initiate time stamp and assign clerk for service. Service will store in service repository it create separate thread for each process and manage service, if service not completed within time stamp then service will suspend for some time based on service class decision. When service class is available then service will resumed.

3.4 Motivation

The Service Administration design pattern decouples the implementation of services from the time at which the services are configured into an application or a system. This decoupling improves the modularity of services and allows the services to evolve over time independently of configuration issues (such as whether two services must be co-located or what concurrency model will be used to execute the services). In addition, the Service Administration design pattern centralizes the administration of the services it configures. This facilitates automatic initialization and termination of services and can improve performance by factoring common service initialization and termination patterns into efficient reusable components.

3.5 Proposed Design Pattern Structure

A UML class diagram for the Service Administration Pattern can be found in Figure 2.

Three design patterns are used for service administration design pattern that are Reflective monitoring, strategy and thread per connection pattern. Client invoke service in service class, observer will observer invocation in service class then report invocation to service class based on the service class will choose appropriate time stamp and clerk for service. After assigning time stamp service stored in service repository. Service repository will handle service based on thread per connection pattern; Pattern will create separate thread for each and every service. Each service is executed in separate location clerk will observe service weather it finishes with in time or not, if service finishes with in time stamp then clerk will report results to client otherwise it report observation to service class. Service takes decision based on availability of time stamp if time is available then service is refused otherwise service is suspended [5]. Here three design patterns are satisfies all autonomic properties monitoring, decision making, reconfiguration and execution. Monitoring reflection monitoring is used for decision making strategy is used, executing thread per connection is used finally service class is reconfigured based on service results either refuse or suspend. Figure 3 will shows sequence diagram for service administration.

3.6 Participants

(a) **Client**
Client class will invoke service in service class, client provide input to service administration system client will try invoke service that are there in service class, if service is there then it will invoke otherwise it get error message if service is there in service class after execution of service client will get result from service class [13].
(b) **Service**
Service class will consists of service it is observed by observer class it will report based on invocation, based on observer it will allocate times stamp and clerk to service. After assign time stamp and clerk service is stored in service repository. Based on service result service class is reconfigured.

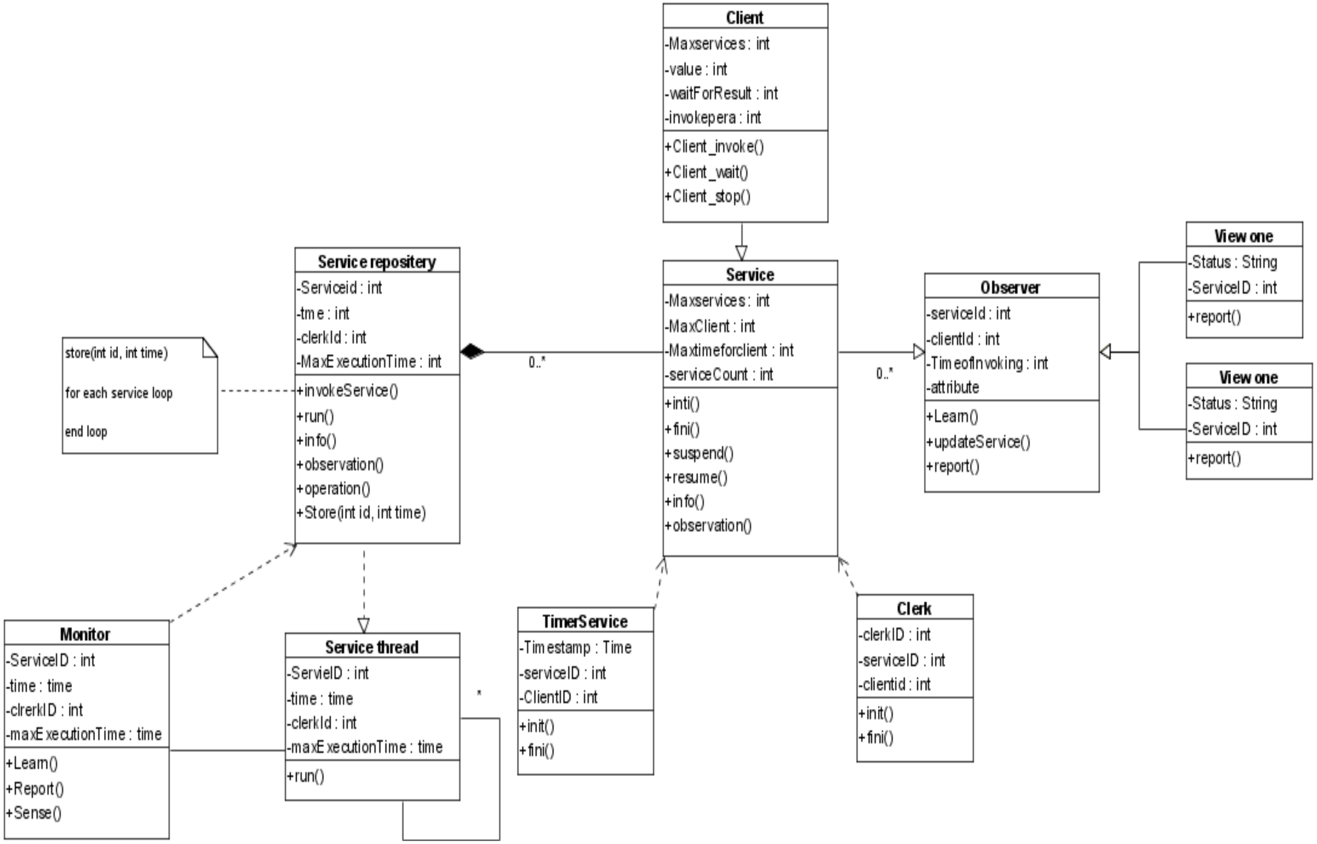


Fig. 2. Class Diagram for Service Configuration

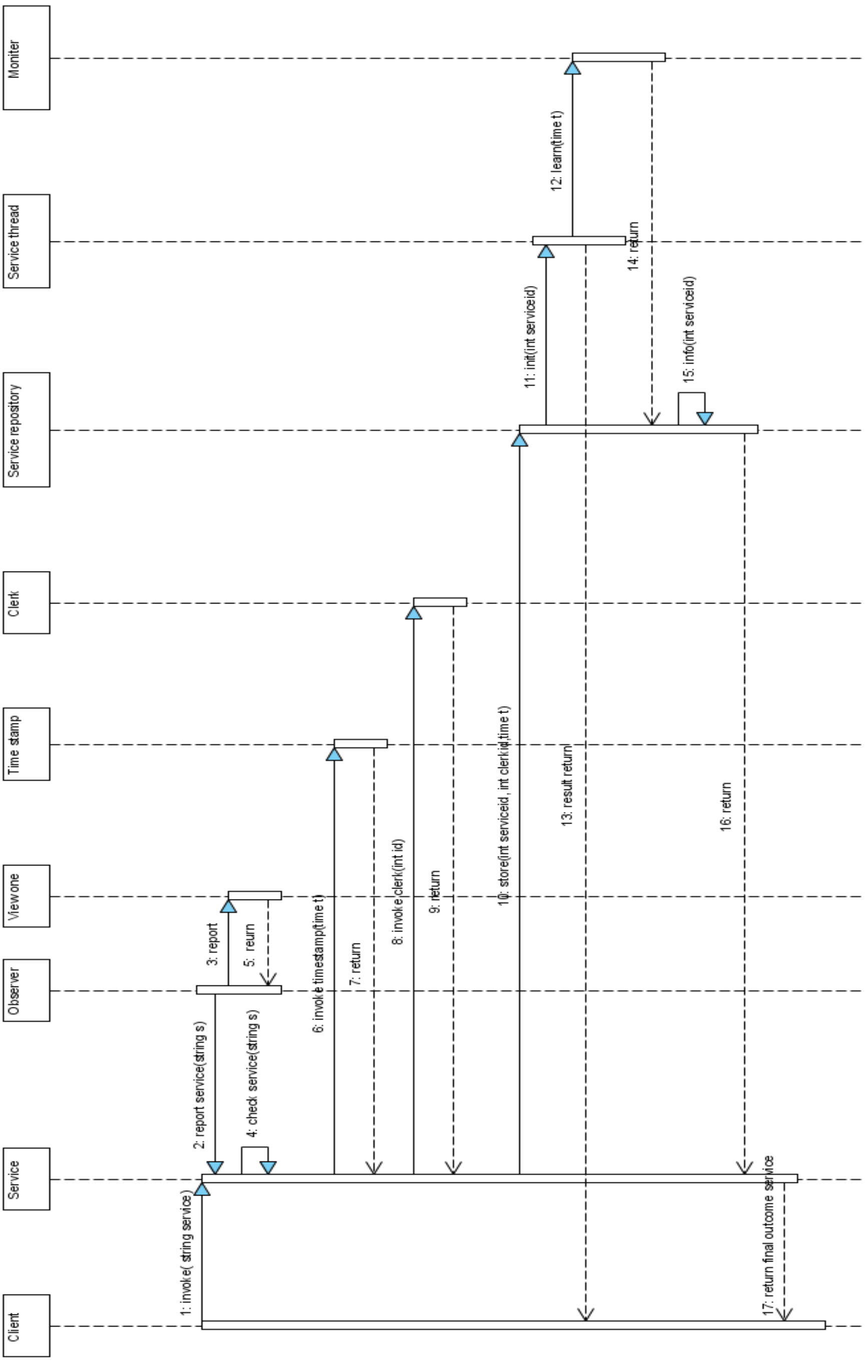


Fig. 3. Sequence Diagram for Service Administration design pattern

(c) **Service Repository**

Service repository stores services that are invoked by client, repository will create separate thread for each client, if service is finished with in time stamp then it will report result to client otherwise it report service class, service class will take decision based on time stamp availability if time is available then it choose resume service otherwise they choose suspend service.

(d) **Time Stamp**

The Time stamp uses the Acceptor class to accept connections from one or more Clerks. The Acceptor class uses the Acceptor pattern to create handlers for every connection from Clerks that want to receive requests for time updates. This design decouples the implementation of the Time Stamp from its configuration. Therefore, developers can change the implementation of the Time Stamp independently of its configuration. This provides flexibility with respect to evolving the implementation of the Time stamp.

(e)**Clerk**

The Clerk uses the Connector class to establish and maintain connections with one or more Time stamp. The Connector class uses the Connector pattern to create handlers for each connection to a Time Server. The handlers receive and process time updates from the Time stamp. The Clerk class inherits from the Service base class.

(f)**Observer**

Observer will monitor service class for service invocation is service in invoked then observer will report service class about service invoking. Based on observer service class choose time stamp and clerk [13].

(f) **Service Thread**

Service is executed in service thread in service thread clerk observes weather service is completed within time stamp or not, Based on result service class is reconfigured.

Service Administration to configure and control a service:

Service initialization - provide an entry point into the service and perform initialization of the service.

Service termination - terminate execution of a service.

Service suspension - temporarily suspend the execution of a service.

Service resumption - resume execution of a suspended service.

Service information- report information (*e.g.,* port number or service name) that describes a service.

3.7 Consequences

The Service Administration design pattern offers the following benefits:

Centralized Administration: The pattern consolidates one or more services into a single administrative unit. This helps to simplify development by automatically performing common service initialization and termination activities. In addition, it centralizes the administration of communication services by imposing a uniform set of configuration management operations.

Increased Modularity and Reuse: The pattern improves the modularity and reusability of communication services by decoupling the implementation of these services from the configuration of the services. In addition, all services have a uniform interface by which they are configured, thereby encouraging reuse and simplifying development of subsequent services.

Increased Configuration Dynamism: The pattern enables a service to be dynamically reconfigured without modifying, recompiling, or statically relining existing code. In addition, reconfiguration of a service can often be performed without restarting the service or other active services with which it is co-located.

Increased Opportunity for Tuning and Optimization: The pattern increases the range of service configuration alternatives available to developers by decoupling service functionality from the concurrency strategy used to execute the service. Developers can adaptively tune daemon concurrency levels to match client demands and available OS processing resources by choosing from a range of concurrency strategies. Some alternatives include spawning a thread or process upon the arrival of a client request or pre spawning a thread or process at service creation time.

3.8 Related Design Patterns

The intent of the Service Administration pattern is similar to the Configuration pattern. The Configuration pattern decouples structural issues related to configuring services in distributed applications from the execution of the services themselves. The **Configuration pattern [1]** has been used in frameworks for configuring distributed systems to support the construction of a distributed system from a set of components. In a similar way, the Service Administration design pattern decouples service initialization from service processing. The primary difference is that the Configuration pattern focuses more on the active composition of a chain of related services, whereas the Service Administration design pattern focuses on the dynamic initialization of service handlers at a particular endpoint. In addition, the Service Administration design pattern focuses on decoupling service behavior from the service's concurrency strategies.

The **Manager Pattern [7]** manages a collection of objects by assuming responsibility for creating and deleting these objects. In addition, it provides an interface to allow clients access to the objects it manages. The Service Administration design pattern can use the Manager pattern to create and delete Services as needed, as well as to maintain a repository of the Services it creates using the Manager Pattern. However, the functionality of dynamically configuring, initializing, suspending, resuming, and terminating a Service created using the Manager Pattern must be added to fully implement the Service Administration Pattern.

3.9 Applicability

Use the Service Administration design pattern when:

- Services must be initiated, suspended, resumed, and terminated dynamically; And
- An application or system can be simplified by being composed of multiple Independently developed and dynamically configurable services; or
- The management of multiple services can be simplified or optimized by Configuring them using single administrative unit.

4 Interfaces Definition for the Design Pattern Entities

(a) **Client:**

Client.java

```
Public class Client
{
   Public invoke( int serviced){}
 }
```

(b) **Service:**

Service.java

```
Public class service
{
 Public init(int serviced){}
 Public observe( obj){}
 Public suspend(int serviced, int
clerkID){}
 Public refuse(int serviced, int
clerkID){ }
 Public fini(int serviced, int clerkID)
{ }
}
```

(c) **Service repository:**

Serviecerepository.java

```
Public class Serviecerepository
{
 Public Invokeservice(int ser){ }
 Public run(int ser, int clerk ){}
 Public store(int serID, int time){ }
}
```

d) **Time stamp:**

Timestamp.java

```
Public class Timestamp
{
 Public init(int serviceID){}
 Public fini(int serviceID){}
}
```

(e)**Clerk:**

Clerk.java

```
Public class Clerk
{
 Public init(int serviceID){}
 Public fini(int serviceID){}
}
```

(f)**Observer:**

Observer.java

```
Public class Observer
{
  Public Leran(){}
  Public Update(){}
  Public Report(){}
}
```

(f) **Service thread:**

Servicethread.java

```
Public Servicethread
{
  Public run()
  {
    }
}
```

5 Profiling Results

To demonstrate the efficiency of the pattern we took the profiling values using the Netbeans IDE and plotted a graph that shows the profiling statistics when the pattern is applied and when pattern is not applied. This is shown in figure 4. Here X-axis represents the runs and Y-axis represents the time intervals in milliseconds. Below simulation shows the graphs based on the performance of the system if the pattern is applied then the system performance is high as compared to the pattern is not applied.

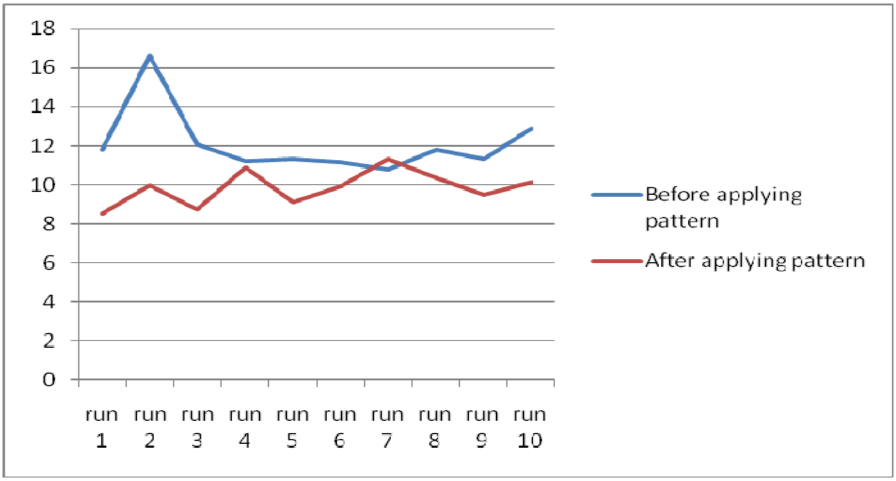


Fig. 4. Profiling statistics before applying pattern and after applying pattern

6 Conclusion

This paper describes the Service Administration design pattern and illustrates how it decouples the implementation of services from their configuration. This decoupling increases the flexibility and extensibility of services. In particular, service implementations can be developed and evolved over time independently of many issues related to service administration. In addition, the Service Administration design pattern provides the ability to reconfigure a service without modifying, recompiling, or statically linking existing code. Based on proposed system existing system will upgrade their resources and their services. Proposed system also satisfies all properties of autonomic system like monitoring, decision making, and reconfiguration. Out future aim is to implement this paper in aspect oriented programming that satisfies all autonomic characteristics of autonomic system.

References

1. Jain, P., Schmidt, D.C.: Service Configuration: A Pattern for Dynamic Configuration of Services. In: 3rd USENIX Conference on Object-Oriented Technologies and Systems Portland (1997)
2. Ramirez, A.J., Betty, H.C.: Design patterns for developing dynamically adaptive Systems. In: 5th International Workshop on Software Engineering for Adaptive and Self-Managing Systems, Cape Town, South Africa, pp. 29–67, 50, 68 (2010)
3. Mannava, V., Ramesh, T.: A Novel Event Based Autonomic Design Pattern for Management of Webservices. In: Wyld, D.C., Wozniak, M., Chaki, N., Meghanathan, N., Nagamalai, D. (eds.) ACITY 2011. CCIS, vol. 198, pp. 142–151. Springer, Heidelberg (2011)
4. Prasad Vasireddy, V.S., Mannava, V., Ramesh, T.: A Novel Autonomic Design Pattern for Invocation of Services. In: Wyld, D.C., Wozniak, M., Chaki, N., Meghanathan, N., Nagamalai, D. (eds.) CNSA 2011. CCIS, vol. 196, pp. 545–551. Springer, Heidelberg (2011)

5. Gamma, E., Helm, R., Johnson, R., Vlissides, J.: Design Patterns Elements of Reusable Object-Oriented Software, Hawthorne, New York (1997)
6. Pree, W.: Design Patterns for Object-Oriented Software Development. Addison-Wesley, MA (1994)
7. Schmidt, D.C., Suda, T.: An Object-Oriented Framework for Dynamically Configuring Extensible Distributed Communication Systems. In: IEE/BCS Distributed Systems Engineering Journal (Special Issue on Configurable Distributed Systems), pp. 280–293 (1994)
8. Crane, S., Magee, J., Pryce, N.: Design Patterns for Binding in Distributed Systems. In: The OOPSLA 1995 Workshop on Design Patterns for Concurrent, Parallel, and Distributed Object-Oriented Systems, Austin, TX. ACM (1995)
9. Chawla, A., Orso, A.: A generic instrumentation framework for collecting dynamic information. SIGSOFT Softw. Eng. Notes (2004)
10. Cheng, S.-W., Garlan, D., Schmer, B.: Architecture-based selfadaptation in the presence of multiple objectives. In: International Workshop on Self-Adaptation and Self-Managing Systems. ACM, New York (2006)

Chemical Process Fault Diagnosis Based on Sensor Validation Approach

Jialin Liu

Department of Information Management, Fortune Institute of Technology,
1-10, Nwongchang Rd., Neighborhood 28, Lyouciyou Village,
Daliao Township, Kaohsiung, Taiwan, Republic of China
jialin@center.fotech.edu.tw

Abstract. Investigating the root causes of abnormal events is a crucial task for an industrial chemical process. When process faults are detected, isolating the faulty variables provides additional information for investigating the root causes of the faults. Numerous data-driven approaches require the datasets of known faults, which may not exist for some industrial processes, to isolate the faulty variables. The contribution plot is a popular tool to isolate faulty variables without a priori knowledge. However, it is well known that this approach suffers from the smearing effect, which may mislead the faulty variables of the detected faults. In the presented work, a contribution plot without the smearing effect to non-faulty variables was derived. An industrial example, correctly isolating faulty variables and diagnosing the root causes of the faults for the compression process, was provided to demonstrate the effectiveness of the proposed approach for industrial processes.

Keywords: Fault detection and diagnosis, Principal component analysis, Contribution plots, Missing data analysis.

1 Introduction

In modern chemical processes, distributed control systems are equipped for regulating the processes, and the operating data are collected and stored in a historical database. However, information about process operations is hidden under the historical data. Therefore, it is more practical to develop methods that detect and investigate the root causes of process faults based on data-driven approaches, rather than to use other methods based on rigorous process models or knowledge-based approaches. Since the measured variables are correlated for a chemical process, principal component analysis (PCA) is a popular tool to extract the features of the process data that are applied to monitor the process variations. After a fault is detected, the faulty variables need to be isolated in order to diagnose the root causes of the fault. Contribution plots are the most popular tool for identifying which variables are pushing the statistics out of their control limits. Kourti and MacGregor [1] applied the contribution plots of quality variables and process variables to find faulty variables of a high-pressure low-density polyethylene reactor. They remarked that the contribution plots may not reveal the assignable causes of abnormal events; however, the group of variables

J.-S. Pan, S.-M. Chen, N.T. Nguyen (Eds.): ACIIDS 2012, Part I, LNAI 7196, pp. 64–73, 2012.

contributed to the detected events will be unveiled for further investigation. Westerhuis et al. [2] introduced the confidence limits of the contribution plots to enhance the capability of identifying the behaviors of faulty variables departing from the normal operating condition (NOC). They reported that there must be a careful interpretation of the contribution plots, since the residuals of the PCA are smeared out over the other variables. Yoon and MacGregor [3] comprehensively compared model-based and data-driven approaches for fault detection and isolation, and summarized that the contribution plots provide for the easy isolation of simple faults, but that additional information about operating the process is needed to isolate complex faults.

Other than isolating faulty variables with contribution plots, fault isolation approaches have been conducted based on different groups of operating data. Raich and Çinar [4] built several PCA models using normal and abnormal process data. The detected faults are diagnosed by comparing the statistical distances and angles of the new data with each group of known event data. Dunia and Qin [5] developed the reconstruction-based approach to isolate faulty variables from the subspaces of faults. Yue and Qin [6] combined the statistics Q and T^2 of PCA to develop an index that is minimized when isolating the faulty variables; therefore, a more feasible solution could be found than that from the original approach. The reconstruction-based contribution (RBC) [7] approach has been derived recently, and it was reported that RBC will not suffer the smearing effect, as the contribution plots of the PCA are enduring. In reality, the smearing effect of RBC can be observed when implementing the confidence intervals of the RBC plots. However, this type of approach, which constructs fault isolation models from the known event data, will induce an incorrect result when encountering a new fault.

In order to identify the faulty variables for a new process fault, He et al. [8] used k-means clustering to classify historical data into different groups. The pairwise Fisher discriminant analysis (FDA) was then applied to the normal data and each class of faulty data to find fault directions that were used to generate contribution plots for isolating faulty variables. Since their approach is only concerned with the variable directions between the classes of normal and faulty data, different classes of faults may have the same faulty variables, when the faulty classes spread in the same directions at different locations. Liu and Chen [9] used Bayesian classification to extract multiple operating regions from historical data. A fault identification index was derived based on the dissimilarities between normal and abnormal cluster centers and covariances. The faulty variables of new faults can be isolated by comparing the indices of the measured variables. However, isolating faulty variables by comparing the dissimilarities between normal and abnormal classes is based on a restrictive assumption that the faulty data can be formed into groups. Kariwalaa et al. [10] integrated the branch and bound (BAB) method with the missing variable approach of probabilistic PCA (PPCA) to locate faulty variables. The concept of the approach is similar to the reconstruction-based method, but the known event datasets are not needed. Since the BAB method searches for faulty variables by minimizing the monitoring statistic of PPCA, it can be expected that the solutions of the faulty variables will be inconsistent when the fault is propagating or when the controllers try to bring the process back to NOC. In the presented work, a contribution plot without smearing effect was derived. In this approach, it is not necessary to prepare the known event datasets and the time-consuming task of continuously optimizing the mixed-integer programming problem for every sampling data until reaching a stable solution is also not required.

The remainder of this paper is organized as follows. Section 2 gives an overview of PCA and the contribution plots of statistics Q and T^2. The proposed approach of the contribution plots without smearing effect to non-faulty variables is detailed in section 3. In section 4, an industrial application is provided to demonstrate the effectiveness of the proposed approach for industrial processes. Finally, conclusions are given.

2 Basic Theory

2.1 Principal Component Analysis

Consider the data matrix $\mathbf{X} \in R^{m \times n}$ with m rows of observations and n columns of variables. Each column is normalized to zero mean and unit variance. The covariance of the reference data can be estimated as:

$$\mathbf{S} \approx \frac{1}{(m-1)} \mathbf{X}^{\mathrm{T}} \mathbf{X} = \mathbf{P} \mathbf{\Lambda} \mathbf{P}^{\mathrm{T}} + \tilde{\mathbf{P}} \tilde{\mathbf{\Lambda}} \tilde{\mathbf{P}}^{\mathrm{T}} \tag{1}$$

where $\mathbf{\Lambda}$ is a diagonal matrix with the first K terms of the significant eigenvalues and $\mathbf{P}$ contains the respective eigenvectors. The $\tilde{\mathbf{\Lambda}}$ and $\tilde{\mathbf{P}}$ are the residual eigenvalues and eigenvectors, respectively. The data matrix $\mathbf{X}$ can be decomposed as: $\mathbf{X} = \mathbf{X}\mathbf{P}\mathbf{P}^{\mathrm{T}} + \mathbf{X}\tilde{\mathbf{P}}\tilde{\mathbf{P}}^{\mathrm{T}} = \hat{\mathbf{X}} + \mathbf{E}$ with $\hat{\mathbf{X}}$ being the projection of the data matrix $\mathbf{X}$ onto the subspace formed by the first K eigenvectors, named the principal component (PC) subspace, and $\mathbf{E}$ being the remainder of $\mathbf{X}$ that is orthogonal to the subspace.

Statistic Q is defined as a measure of the variations of the residual parts of data: $Q = (\mathbf{x} - \hat{\mathbf{x}})(\mathbf{x} - \hat{\mathbf{x}})^{\mathrm{T}} = \mathbf{x}\tilde{\mathbf{P}}\tilde{\mathbf{P}}^{\mathrm{T}}\mathbf{x}^{\mathrm{T}} = \mathbf{x}\mathbf{C}\mathbf{x}^{\mathrm{T}}$ where $\mathbf{C} \equiv \tilde{\mathbf{P}}\tilde{\mathbf{P}}^{\mathrm{T}}$. In addition, another measure for the variations of systematic parts of the PC subspace is the statistic T^2: $T^2 = \mathbf{x}\mathbf{P}\mathbf{\Lambda}^{-1}\mathbf{P}^{\mathrm{T}}\mathbf{x}^{\mathrm{T}} = \mathbf{x}\mathbf{D}\mathbf{x}^{\mathrm{T}} = \mathbf{t}\mathbf{\Lambda}^{-1}\mathbf{t}^{\mathrm{T}}$ where $\mathbf{D} \equiv \mathbf{P}\mathbf{\Lambda}^{-1}\mathbf{P}^{\mathrm{T}}$ and $\mathbf{t}$ are the first K term scores. This is the Mahalanobis distance from the origin of the subspace to the projection of the data. The confidence limits of Q and T^2 can be found in reference [11]. When a fault is detected by any one of above-mentioned statistics, the contribution plots provide a preliminary tool to isolate faulty variables without any prior knowledge of the fault. The contributions of Q for the i^{th} variable can be written as: $c_i^Q = (\mathbf{x}\mathbf{C}\xi_i)^2$ where ξ_i is a column vector in which the i^{th} element is one and the others are zero. The confidence limit for each contribution of Q has been derived in references [2]. Qin et al. [12] derived the variable contributions to T^2 for the i^{th} variable as $(\mathbf{x}\mathbf{D}^{0.5}\xi_i)^2$, and also provided the confidence limits of the contributions. However, since the contributions of the statistics are transformed from the process variables through a matrix multiplication, the faulty variables may smear out over the other variables, which will mislead a diagnosis of the correct root causes of the faults.

2.2 Fault Isolation Based on Missing Data Approach

Qin et al. [13] developed a sensor validation approach based on the reconstructed missing data. Each variable is validated in order to minimize the statistic Q when it is over its control limit.

$$\left.\left(\frac{\partial Q}{\partial x_k}\right)\right|_{x_i \neq x_k} = 2\sum_{i=1}^{n} x_i c_{k,i} = 0, \quad k = 1...n \tag{2}$$

where the $c_{i,j}$ is an element of the $\mathbf{C}$. Rearranging the above equation, the reconstructed input can be obtained: $x_k^* = -\frac{1}{c_{k,k}}\sum_{i\neq k}^{n} x_i c_{k,i}$ where x_k^* is the k^{th} reconstructed input. Substituting the reconstructed input into the definition of Q, the statistic Q with the k^{th} reconstructed variable can be written as: $Q_k^* = \sum_{i\neq k}^{n}\sum_{j\neq k}^{n} x_i x_j c_{i,j} + 2x_k^*\sum_{j\neq k}^{n} x_j c_{j,k} + x_k^* x_k^* c_{k,k}$. The faulty variable was located by selecting the variable with the minimum of sensor validity indices (SVIs), which was defined as $\eta_k = Q_k^* / Q$ for the k^{th} variable [13].

Yue and Qin [6] suggested that T^2 may be over its control limit when Q is minimized to reconstruct the faulty sensor. Therefore, they proposed a combined index as: $\varphi \equiv \frac{Q}{Q_\alpha} + \frac{T^2}{T_\alpha^2} = \mathbf{x\Phi x}^{\mathrm{T}}$, $\mathbf{\Phi} \equiv \mathbf{C}/Q_\alpha + \mathbf{D}/T_\alpha^2$ where Q_α and T_α^2 respectively are the (1-α) confidence limits of the statistic Q and T^2. The combined index is minimized instead of the Q statistic and the reconstructed variable can be obtained as: $\left.\left(\frac{\partial \varphi}{\partial x_k}\right)\right|_{x_i \neq x_k} = 0, \quad k = 1...n$. The combined index with the k^{th} reconstructed variable can be written as: $\varphi_k^* = \sum_{i\neq k}^{n}\sum_{j\neq k}^{n} x_i x_j \phi_{i,j} + 2x_k^*\sum_{j\neq k}^{n} x_j \phi_{j,k} + x_k^* x_k^* \phi_{k,k}$.

3 Isolating Multiple Sensor Faults

For a multiple sensor fault, the reconstructed variables can be written as: $\boldsymbol{\xi}^{\mathrm{T}}\mathbf{\Phi x}^{\mathrm{T}} = \mathbf{0}$ where $\boldsymbol{\xi} \equiv \begin{bmatrix}\xi_1 & \xi_2 & \cdots & \xi_{nf}\end{bmatrix}$, in which nf is the number of faulty variables and ξ_i is a column vector in which the i^{th} element is one and the others are zero. The monitored variables can be decomposed as $\mathbf{x} = \mathbf{x}\boldsymbol{\eta} + \mathbf{x}(\mathbf{I} - \boldsymbol{\eta})$, where $\boldsymbol{\eta}$ is a diagonal matrix, in which the values of the diagonal elements are one for the faulty variables and zero for the non-faulty ones. Therefore, minimizing the combined index can be rewritten in the following form: $\boldsymbol{\xi}^{\mathrm{T}}\mathbf{\Phi}\boldsymbol{\eta}\mathbf{x}^{\mathrm{T}} = -\boldsymbol{\xi}^{\mathrm{T}}\mathbf{\Phi}(\mathbf{I} - \boldsymbol{\eta})\mathbf{x}^{\mathrm{T}}$. The left-hand side of the equation contains the data to be reconstructed by the normal data in the right-hand term. The data that to be reconstructed can be expressed as: $\boldsymbol{\eta}\mathbf{x}^{\mathrm{T}} = \boldsymbol{\xi}\mathbf{x}_{nf}^{\mathrm{T}}$ in which $\mathbf{x}_{nf}$ is the collection of the faulty variables. The reconstruction of the faulty variables can be obtained by the following equation:

$$\mathbf{x}_{nf}^{*\mathrm{T}} = -\left(\boldsymbol{\xi}^{\mathrm{T}}\boldsymbol{\Phi}\boldsymbol{\xi}\right)^{-1}\boldsymbol{\xi}\boldsymbol{\Phi}\left(\mathbf{I}-\boldsymbol{\eta}\right)\mathbf{x}^{\mathrm{T}} \tag{3}$$

Since $\boldsymbol{\Phi}$ is a full-rank matrix, $\left(\boldsymbol{\xi}^{\mathrm{T}}\boldsymbol{\Phi}\boldsymbol{\xi}\right)$ is invertible. The reduction of the combined index (RCI) after the reconstructing the faulty data can be written as follows:

$$\varphi - \varphi_{nf}^{*} = \left(\mathbf{x}_{nf} - \mathbf{x}_{nf}^{*}\right)\left(\boldsymbol{\xi}^{\mathrm{T}}\boldsymbol{\Phi}\boldsymbol{\xi}\right)\left(\mathbf{x}_{nf} - \mathbf{x}_{nf}^{*}\right)^{\mathrm{T}} \tag{4}$$

Therefore, the fault isolation task is to find subset $\mathbf{x}_{nf}$ from $\mathbf{x}$ to maximize the RCI, until the statistics Q and T^2 are under the corresponding control limits, without the information from faulty variables. The contribution of RCI for the i^{th} faulty variable can be defined as:

$$c_i^{RCI} = \left[\left(\mathbf{x}_{nf} - \mathbf{x}_{nf}^{*}\right)\left(\boldsymbol{\xi}^{\mathrm{T}}\boldsymbol{\Phi}\boldsymbol{\xi}\right)^{0.5}\boldsymbol{\xi}_i\right]^2 \tag{5}$$

The proposed approach first evaluates each RCI with a reconstructed variable and inserts the variable with the maximum RCI into $\mathbf{x}_{nf}$ in the first step. Next, the RCIs are evaluated using the reconstructed data of a non-faulty variable and the selected faulty variables in $\mathbf{x}_{nf}$. The non-faulty variable with the maximum RCI is inserted into $\mathbf{x}_{nf}$. The steps of adding a new faulty variable into $\mathbf{x}_{nf}$ is repeated until both statistics are under the corresponding control limits. The algorithm is summarized as follows:

1. Set $nf = 0$ and $\mathbf{x}_{nf} = \varnothing$.
2. For $i = 1 \ldots n\text{-}nf$,

   Reconstruct the data of $\mathbf{x}_{nf} \cup \mathbf{x}_i \notin \mathbf{x}_{nf}$ using eq. 3 and evaluate the RCI using eq. 4.
3. Add the variable with the maximum RCI into $\mathbf{x}_{nf}$ and set $nf = nf + 1$.
4. If the statistics Q and T^2, without the information of the selected faulty variables, are still over their control limits, go back to step 2.
5. Decreasingly sort the selected faulty variables according to the contributions of RCI using eq. 5 and retain the variables in $\mathbf{x}_{nf}$ that sufficiently reduce the statistics Q and T^2 under the corresponding control limits.

Steps 2-4 of the algorithm guarantee eq. 4 to be a monotonically increasing function with the number of selected faulty variables; therefore, the statistics monotonically decrease during iterations. The non-faulty variables, which may be selected in the early stage of the iterations under insufficient information about the faulty variables, are removed from $\mathbf{x}_{nf}$ in step 5. When diagnosing the root causes of process faults, the selected faulty variables do not equally contribute to the faults. The contribution plots for the reduction of statistics can be used to find the faulty variables with the most contributions, as the contributions have been confined within the selected faulty variables and the fault magnitude will not smear over to the non-faulty variables.

4 Industrial Application

The compression process was a 4-stage centrifugal compressor, equipped with an intercooler between stages to cool down the compressed air, as Fig. 1 shows. In order to reduce the shaft work, the compression ratio for each stage is maintained within a similar range, whereas the inlet temperature of each stage needs to be kept as low as possible. A detailed description of the process can be found in the previous study [9]. PCA was applied to the training dataset, in which the data of measured variables listed in Table 1 were collected for five days, and three PCs were retained by cross-validation, which captured about 87% of the total variance. The process was monitored using the fast moving window PCA (FMWPCA) approach [14] with a five-day window size. The fault detection result is shown in Fig. 2, in which the abnormal events detected by the statistic T^2 around day 1, 2.5 and 8.5 were due to the scheduled maintenances of the process equipments that could be disclosed from the operator log. However, the root causes of the process fault detected by the statistic Q after approximately the sixth day could not be unveiled by the log.

Table 1. Measured variables for the compression process

Measured Variable	Description
F_a (x_1)	Feed flow rate of air
$P_{in,\,i}$ ($x_2 - x_5$)	Inlet pressure for the i^{th} compression stage, i = 1…4
$P_{out,\,i}$ ($x_6 - x_9$)	Outlet pressure for the i^{th} compression stage, i = 1…4
$T_{in,\,i}$ ($x_{10} - x_{13}$)	Inlet temperature for the i^{th} compression stage, i = 1…4
$T_{out,\,i}$ ($x_{14} - x_{17}$)	Outlet temperature for the i^{th} compression stage, i = 1…4
T_c (x_{18})	Inlet temperature of the cooling water
$T_{c,\,i}$ ($x_{19} - x_{21}$)	Outlet temperature of the i^{th} intercooler, i = 1…3

The statistics after removing the faulty variables (FVs) is also shown in Fig. 2. It demonstrates that the proposed approach guarantees the statistic Q and T^2 under their control limits after removing the FVs. The normalized contribution plot of RCI is shown in Fig. 3(a) that indicates the major faulty variables after the sixth day was x_{15}, which is the outlet temperature of the second compression stage. The fault isolation results using the contribution plot of Q is displayed in Fig. 3(b), in which each contribution was normalized with the corresponding 99% confidence limits. Although Fig. 3(b) indicates that x_{15} was one of the most significant faulty variables; however, the fault magnitude was smeared over to the non-faulty variables. Comparing the results of Fig. 3(a) and 3(b), the smearing effect of the traditional contribution plots is effectively eliminated using the proposed approach. Figure 3(c) shows the RBC of Q normalized with the corresponding 99% confidence limits that is exactly same with Fig. 3(b). Since the RBC of Q and the traditional contribution plots of Q differ only by a scaling coefficient, which appears in the leading part of the control limits of RBC, the normalized RBC would be identical to the normalized contribution plots.

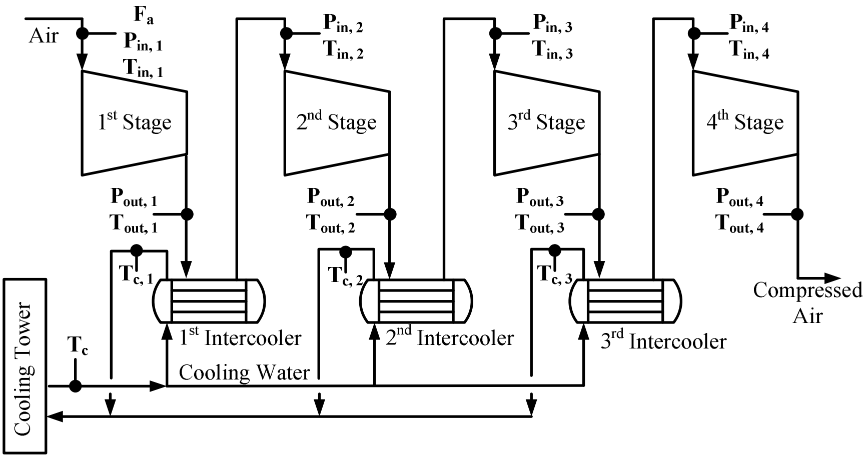


Fig. 1. Air compression process flow diagram

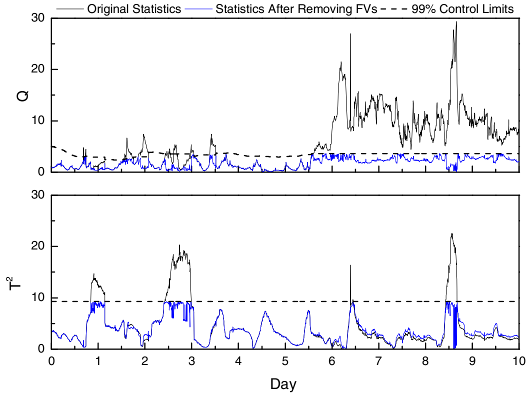
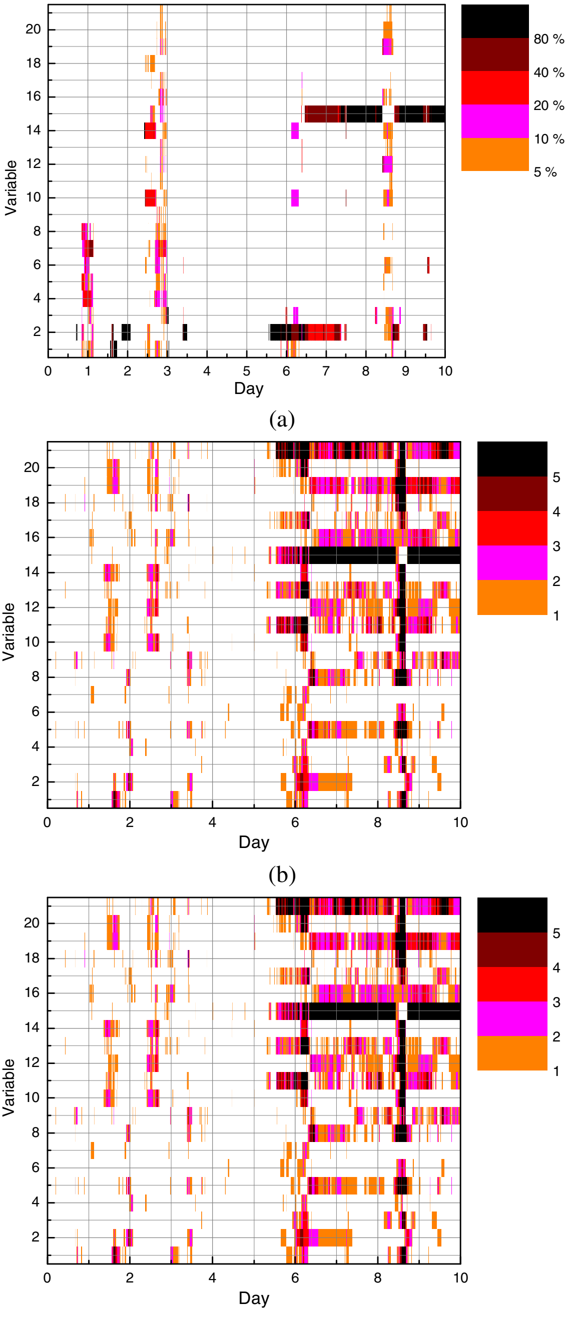


Fig. 2. Process monitoring using FMWPCA

For diagnosing root causes of the fault after the sixth day, the measured and reconstructed data of x_{15} are shown in Fig. 4(a) in which the measurements were lower than the reconstructed data during the period that the fault was detected. Comparing these lower temperature data with the normal operating data, it can be found that the variations of faulty data were still under the range of normal operation. Therefore, the statistic Q detected the fault due to the variable correlation changes. For each stage of a centrifugal compressor, the compression efficiency, which can be evaluated from the inlet-outlet temperatures and pressures, is an important index to evaluate the operating performance of the compression stage. Since the compression efficiency is reversely proportional to the discharge temperature of the stage, it can be expected that the faulty data of x_{15}, which is the outlet temperature of the second stage, would mislead the compression efficiency of the second stage being too high. Figure 4(b) compares the compression efficiencies of all stages. The figure shows that the second stage's efficiency extremely fluctuated after the fault had been detected; therefore, it can be concluded that the root cause of the detected fault was the sensor unreliability of the second stage's outlet temperature.

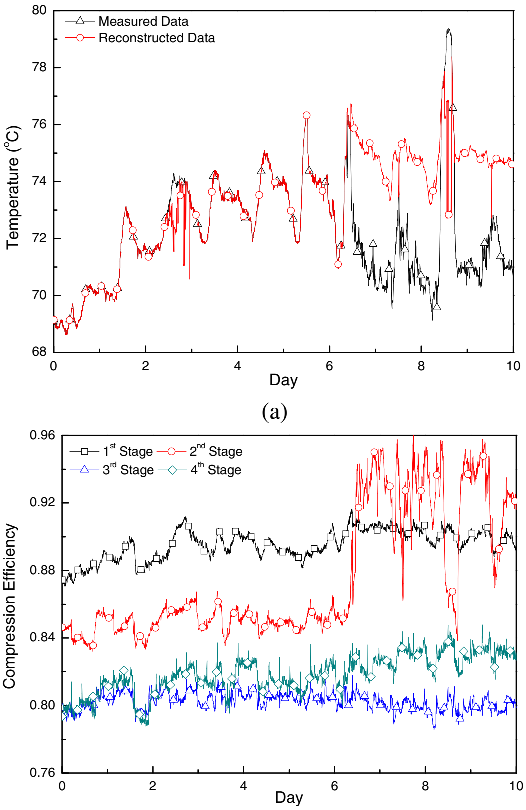


(a)

(b)

(c)

Fig. 3. Comparing the fault isolation results, (a) the proposed approach, (b) contribution plots of Q normalized with the corresponding 99% confidence limits, (c) the RBC of Q normalized with the corresponding 99% confidence limits



(a)

(b)

Fig. 4. Diagnosing root causes of the fault after the sixth day, (a) comparison of the measured and reconstructed data of x_{15}, (b) comparison of compression efficiencies for all stages

5 Conclusions

The presented work developed a contribution plot without the smearing effect to non-faulty variables. The proposed approach was shown to have the capability of isolating multiple sensor faults without predefined faulty datasets. Since the resolution of predefined faulty datasets would be deteriorated due to the time-varying nature of industrial processes, it is not practical to isolate faulty variables based on the historical event lists of an industrial process. In the industrial application, the fault isolation results using the contribution plots of RCI were more precise than the solutions found using the traditional contribution plots. In addition, it was demonstrated that the normalized RBC of Q is equivalent to the traditional contribution of Q normalized with the corresponding control limits; therefore, the RBC approach still suffers the smearing effect when encountering a new fault. The results show that the predefined faulty datasets are not necessary for the proposed approach; in addition, the smearing effect of the traditional contribution plots is also eliminated.

References

1. Kourti, T., MacGregor, J.F.: Multivariate SPC Methods for Process and Product Monitoring. J. Qual. Technol. 28, 409–428 (1996)
2. Westerhuis, J.A., Gurden, S.P., Smilde, A.K.: Generalized Contribution Plots in Multivariate Statistical Process Monitoring. Chemom. Intell. Lab. Syst. 51, 95–114 (2000)
3. Yoon, S., MacGregor, J.F.: Statistical and Causal Model-Based Approaches to Fault Detection and Isolation. AIChE J. 46, 1813–1824 (2000)
4. Raich, A., Çinar, A.: Statistical Process Monitoring and Disturbance Diagnosis in Multivariable Continuous Processes. AIChE J. 42, 995–1009 (1996)
5. Dunia, R., Qin, S.J.: Subspace Approach to Multidimensional Fault Identification and Reconstruction. AIChE J. 44, 1813–1831 (1998)
6. Yue, H.H., Qin, S.J.: Reconstruction-Based Fault Identification Using a Combined Index. Ind. Eng. Chem. Res. 40, 4403–4414 (2001)
7. Alcala, C.F., Qin, S.J.: Reconstruction-based Contribution for Process Monitoring. Automatica 45, 1593–1600 (2009)
8. He, Q.P., Qin, S.J., Wang, J.: A New Fault Diagnosis Method Using Fault Directions in Fisher Discriminant Analysis. AIChE J. 51, 555–571 (2005)
9. Liu, J., Chen, D.S.: Fault Detection and Identification Using Modified Bayesian Classification on PCA Subspace. Ind. Eng. Chem. Res. 48, 3059–3077 (2009)
10. Kariwalaa, V., Odiowei, P.E., Cao, Y., Chen, T.: A Branch and Bound Method for Isolation of Faulty Variables through Missing Variable Analysis. J. Proc. Cont. 20, 1198–1206 (2010)
11. Jackson, J.E.: A User's Guide to Principal Components. Wiley, New York (1991)
12. Qin, J.S., Valle, S., Piovoso, M.J.: On Unifying Multiblock Analysis with Application to Decentralized Process Monitoring. J. Chemom. 15, 715–742 (2001)
13. Qin, S.J., Yue, H., Dunia, R.: Self-Validating Inferential Sensors with Application to Air Emission Monitoring. Ind. Eng. Chem. Res. 36, 1675–1685 (1997)
14. Wang, X., Kruger, U., Irwin, G.W.: Process Monitoring Approach Using Fast Moving Window PCA. Ind. Eng. Chem. Res. 44, 5691–5702 (2005)

System Analysis Techniques in eHealth Systems: A Case Study

Krzysztof Brzostowski, Jarosław Drapała, and Jerzy Świątek

Wrocław University of Technology, The Institute of Computer Science,
Wybrzeże Wyspiańskiego 27, 50-370 Wrocław, Poland
{krzysztof.brzostowski,jaroslaw.drapala,jerzy.swiatek}@pwr.wroc.pl
http://www.pwr.wroc.pl

Abstract. In the paper problem of planning training protocol with taking into account limitations on the training intensity due to the health problems of the exerciser is considered. In the first part of the work short introduction to existing solutions in the area of eHealth applications is given. Next, architecture of the eHealth system to support exerciser training is discussed. The main functionalities of proposed system are pointed out and challenges are highlighted. The concept of context-awareness and personalization is stressed. At the end the problem of model based optimisation of the training protocol is formulated.

Keywords: eHealth, medicine, sport science, modelling, decision suport systems, wireless technologies.

1 Introduction

Putting together wireless sensor network technologies with system analysis techniques in distributed computing environment effects useful and powerful tool capable to put into practise in various areas such as industry, healthcare, sport, emergency management and entertainment. Such systems operating in distributed environment can be composed of huge amount of sensing nodes (e.g. Bluetooth or ZigBee interfaces) with wireless transceivers, computational units i.e. servers and remote nodes to presents results of computing. Moreover these systems must be capable to transfer and process giant volume of data generated by any number of users access nodes. It must be stressed that in many cases (it is conditional upon application area) collected data must be processed in real-time. Additionally, in order to improve quality of usability in user's point of view it is necessary to assure access to systems functionalities anytime and anywhere. It means that in provided systems among other things seamless handover mechanism must be implemented. This mechanism keeps continuous Internet connectivity even sudden change of network occurred and it is implemented in mobile IP protocol. There are two version of mobile IP i.e.: for IPv4 and IPv6. Mobile IP for IPv6 based on its version for IPv4 but many improvements have been made. Therefor the first one is strongly recommended for modern systems based on wireless technologies [8].

J.-S. Pan, S.-M. Chen, N.T. Nguyen (Eds.): ACIIDS 2012, Part I, LNAI 7196, pp. 74–85, 2012.

According to WHO (World Health Organisation) chronic disease such as diabetes, cardiovascular diseases are the leading cause of death. In 2004 an estimated 3.4 million of people died from consequences of high blood sugar. Cardiovascular disease was cause of death for 17.1 million people in 2004. It was 29% of all global deaths [15]. Risk factors for diabetes (Type II) and other chronic disease such as cardiovascular disease are e.g. unhealthy diet and low level of physical activity. They affect decreasing people's health and usually leads to increasing costs of healthcare. Some of these costs can be reduced by delivering tools which are able to promote healthy lifestyle and supporting healthy diet habits.

Recently researchers as well as companies have been grown their efforts towards pervasive systems takes the advantage of wearable sensing devices and wireless sensor networks to elaborate easy-to-use systems to support management and monitoring of healthy lifestyle. Usually such platform are composed of sensing devices operates with Bluetooth and ZigBee interfaces. They are helpful to design sensing platforms that are comfortable to use and not bothersome. Depending on used sensing devices the platforms are called BAN (Body Area Network) or PAN (Personal Area Network). Both BAN and PAN are connected with gateway facilitates transfer sensed data through Internet to server (computational unit) and/or other user's access devices.

2 eHealth Systems: State of the Art

On the market few different commercial equipments to wireless sensing of physiological and kinematic data are produced e.g. by Zephyr Technology and Shimmer-Research manufactures. The first one is a developed construction produced with Bluetooth interface and has ability to sense following physiological signals: heart rate, breath rate, temperature, ECG and others such as acceleration. The second product is open and is still under development but many successful applications has been reported by both researchers and engineers teams [7], [9], [11]. Technical documents offered by Shimmer Research manufacturer can be used to design various sensing systems to be applied in various eHealth areas. Configuration of the platform and applied sensors depends on application. It means that, for example, platform for athletes and its configuration can be different from the platform for monitoring old and impaired people. Shimmer Research offers following sensors: ECG (Electrocardiography), EMG (Electromyography), GSR (Galvanic skin response), Acceleration, Gyroscope, GPS etc.

The second group of products, interesting for engineers and designers of systems to support healthy life style, are devices to measure blood glucose level, blood pressure and body weight. Conducted research of commercial systems shows that many manufacturers of mentioned measuring units offer patients web-based systems to manage their data. Main functionality of such systems is to support transferring data from glucometer and blood pressure meter to remote server by use of wireless technologies. Moreover, patients are allowed to look through historical data. It is possible to make some simple processing of gathered data. In cooperation with physician it may be fixed several kinds of events such as alarms or reminders e.g. about taking medicine. Described web-based systems are offered among others by BodyTel, Entra Health Systems or Lifescan [12],[13],[14]. They also provide glucometers with Bluetooth interface or, like LifeScan, special adapters which allows to add Bluetooth interface to old glucometers without such interface. It is very interesting and well suited solution especially for elderly people.

On the other hand we credited prototypes in which many innovative solutions are proposed. In [6] authors proposed a system which adds to contemporary commercial systems some new functionalities. One of them is connected with supporting a diet management, which is very important in the diabetes therapy. The other functionality is an ability to predict a glucose level after an insulin injection. To this end proposed system takes into account glucose level in blood just before insulin injection and additionally historical data. In [3], [4] and [5] models of glucose-insulin are proposed. It is another step forward enhance knowledge on diabetic.

Some other example of the eHealth system for diabetes treatment is presented in [10]. This system includes controlling an insulin pump. Utilisation of the insulin pump is very interesting and it can lead to very innovative and useful solutions but so far there are many technical problems related to their employment in medical application.

3 Architecture of the eHealth Systems

Wearable sensors to acquire human physiological and kinetic data and portable devices like cellular phones and smart phones became more and more available. Mentioned above company Zephyr Technology offers ready to use equipments in low price. It takes effect that users start to carry these devices on their body and around their body. Because most of these devices have wireless links such as Bluetooth or ZigBee it is possible to connect them in order to create Body Area Networks or Personal Area Networks. Then, sensed data from user and his/her environment can be transferred via the gateway nodes to server. On the server site data are processed and the results are making accessible to mobile client. Mobile client is usually used by supervisor to monitor the process management. In the area of interests i.e. healthcare and wellness applications by the process management we mean the management of treatment or sport training respectively. In the first case the application is designed for physician and in the second one it is intended for sport trainer.

Taking into account mentioned above elements i.e. wireless sensing units, servers and mobile client it is possible to build distributed computational environment for both medical and wellness applications.

Proposed architecture (see Fig. 1) is three-tier architecture composed of data acquisition tier (BAN, PAN); data processing and support for decision making tier and the last one which is presentation tier. The first tier is wireless sensor network that is composed of sensing units that are placed on human body and his/her environment. Each user of the system can use several sensing units.

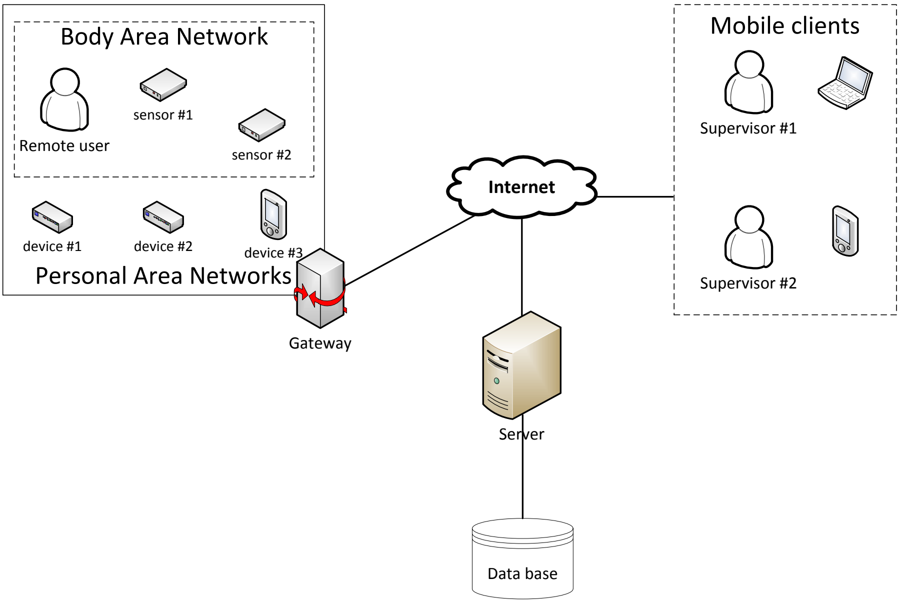


Fig. 1. General system architecture for healthcare and wellness

Sensed data can be pre-processed on personal server such as smart phone or cellular phone and then transferred to server through Internet. It is worth stressing that smart phones and/or cellular phones play two different roles in proposed system. First of them is related to pre-processing of collected data from BAN and PAN. The second one is to provide connection between users wireless sensor networks and remote server through Internet. Personal server that runs on cellular phone or smart phone can be applied only to the tasks which are simple and they are not demanding complex calculations. This unit can be used to manage sensor units consist of BAN or PAN as well.

The central element of the system is second tier with computational units such as servers. On the server site suitable application that provides set of functionalities is running. Architecture of the application allows to configure it and build problem-oriented scenarios. It means that it is possible to choose data processing, supporting decision making and presentation components. These components are used to design well-suited scenarios for certain healthcare or wellness problem. Because of different nature of the sensed signal and problems, different data processing methods may be used. For example, in simple application only signal filtering algorithms are applied. Sophisticated problems may require applied mathematical model. Thus it might be necessary to use estimation algorithms which were usually time consuming. More complex problems, to be solved, in many healthcare or wellness applications are connected with pattern recognition and supporting decision making. In such tasks feature extraction and selection must be solved first, then classification problem can be accomplished. Supporting decision making algorithms typically based on model or/and pattern recognition. They are demanding problems composed of model's parameters estimation, learning and classification algorithms. The another element of the second tier is data base or streaming data base. They are used to store sensed signal, prepared learning sets, estimated model's parameters and results of decision making.

The last tier of the system is presentation tier and its main feature is visualising (e.g. charts) and reporting (e.g. tables) results of data processing and supporting decision making. In Fig. 2 architecture of application has been presented.

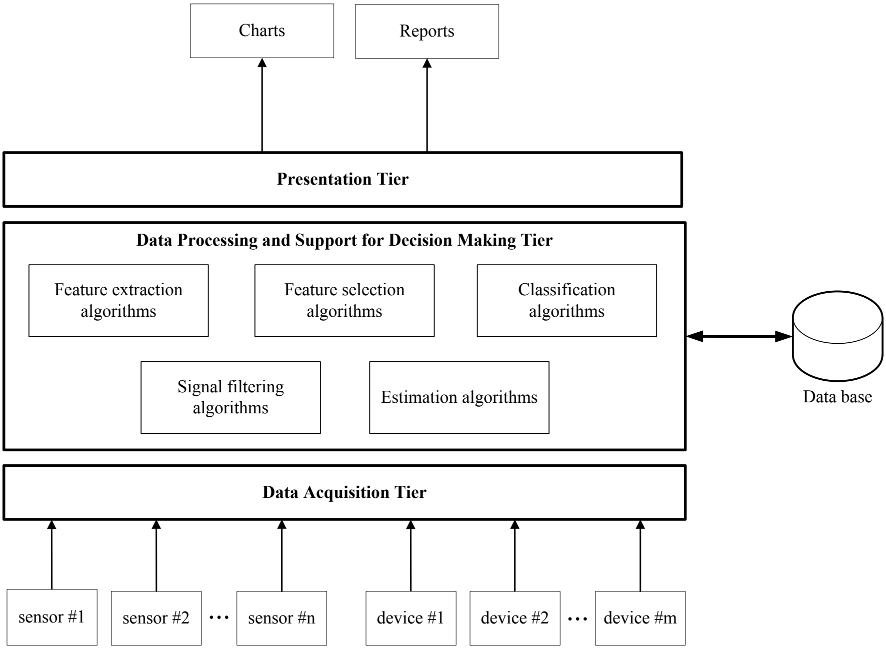

Charts
Reports
Presentation Tier
Data Processing and Support for Decision Making Tier
Feature extraction algorithms
Feature selection algorithms
Classification algorithms
Signal filtering algorithms
Estimation algorithms
Data base
Data Acquisition Tier
sensor #1
sensor #2
sensor #n
device #1
device #2
device #m

Fig. 2. Application architecture for healthcare and wellness

3.1 Proposed Architecture: Benefits Achieved

In the previous section we have stressed that some of the system's feature are simple and some of them are sophisticated. Methods and algorithms constitute the second group of features are more computational demanding. But applied them in our application we had gain new and useful features which would impact on user comfort and improve quality of usability.

One of the profit provided by proposed architecture is context-awareness. Estimating user's context is possible by make use of BAN and PAN and by fusion of data from sensing units. The context information are very useful to understand condition of user, his/her location and environment. It can be helpful for supporting decision making. It means that we have a tool which helps us to improve quality of support for decision making by taking advantage of user's state, recognition of daily activities, his/her behaviour and environment.

The second advantage which is considered in this work and come out of proposed architecture to process data from large amount of sensing units which are placed on user's body and his/her environment is personalisation. In previous section problem of parameter's estimation of model has been discussed. Mathematical model can help us to extract knowledge on considered process or object from acquired data. Then it can be used to predict the future behaviour, activities or actions of the specific user and it helps to improve management of network and computational resources.

4 System Analysis in Healthcare and Wellness: Case Study

Since large volume of measurement data may be sent to computer centers and processed in a real-time, advanced processing algorithms may be performed on the data. Now, we describe the application scenario from the system modeling and analysis viewpoint.

The eHealth system is employed to acquire and deliver data and signals for system identification, optimization and control. Among many functionalities, one of it's role is to support management of dynamic exercise intensity. The following stages of data processing take place:

- measurement data acquisition,
- parameters of the exerciser's model estimation,
- model based optimization of the training protocol (planning),
- training support by the control algorithm.

Depending on the sport discipline, different training goals may be formulated. We focus our attention on footrace, where all stages of data processing mentioned before are present. The goal of footrace is to run a given distance in a shortest time. Moreover, some limitations on the training intensity may be introduced due to the health problems of the exerciser. Typical health problems include: cardiovascular diseases, obesity and diabetics. In case of exerciser having cardiovascular problems or obesity, the heart rate must be monitored to prevent such incidents as fainting, heart attack and brain stroke. In case of diabetic, it is crucial to keep blood glucose level within the normal range, [5]. This task requires the use of glucometer to measure blood glucose level. Proposed eHealth system provides a unified framework for processing data acquired by body sensors of different types. This simplifies the design of algorithms, since all data are available in a unified manner.

Measurement Data Acquisition. In the basic setting, two variables are measured in a real-time: speed and heart rate. Advanced setting includes measurement of blood glucose level before and after the exercise.

Parameters of the Exerciser's Model Estimation. The model describes relation between the heart rate and exercise intensity. Increasing heart rate allows the cardiovascular system to deliver more blood and oxygen to active muscles. The model takes into account short-term and long-term effects of exercises. It has the form of nonlinear set of differential equations, [1]:

$$\begin{aligned} x_1'(t) &= -a_1 x_1(t) + a_2 x_2(t) + a_2 u^2(t) && (1)\\ x_2'(t) &= -a_3 x_2 + \phi(x_1(t)) \qquad , && (2)\\ y(t) &= x_1(t) \end{aligned}$$

where

$$\phi(x_1) = \frac{a_4 x_1}{1 + \exp(a_5 - x_1)}, \tag{3}$$

where x_1 is heart rate change from the rest (resting heart rate), u denotes speed of the exerciser, x_2 may be considered as fatigue, caused by such a factors as: – vasodilation in the active muscles leading to low arterial blood pressure, – accumulations of metabolic byproducts (e.g. lactic acid), – sweeting and hyperventilation. Parameters $a_1, \ldots, a_5$ take nonnegative values.

Values of these parameters are obtained by the estimation procedure, with use of data measured from few experiments. Typical training protocol involves step-like functions $u(t)$ that determine length of the resting (zero speed), the exercise (high speed) and the recovery (walking or resting) periods (see Fig. 3).

For different training protocols, the heart rate profiles are registered and identification algorithm is applied to obtain values of parameters $a_1, \ldots, a_5$ (see Fig. 4). Note the presence of unmeasurable variable $x_2(t)$ within the subsystem modeling fatigue (the block at the bottom in Fig. 4). In order to handle identification of such a closed-loop nonlinear system, numerical optimization algorithm must be employed (in the work [2], where the treadmill is used to control speed, authors employ the Levenberg-Marquardt procedure). It is reasonable to make use of stochastic search methods, such as Simulated Annealing.

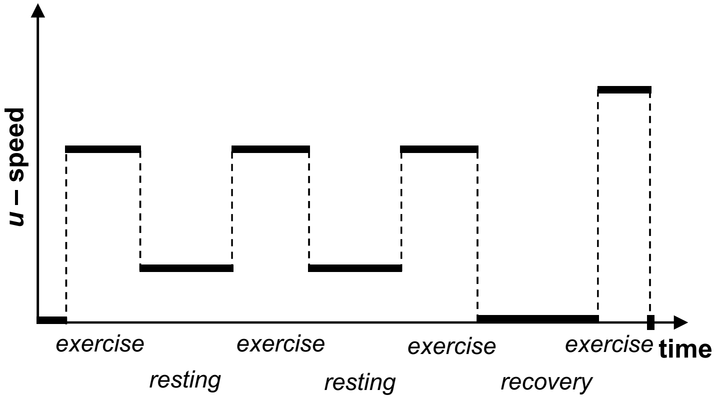


Fig. 3. Typical training protocol

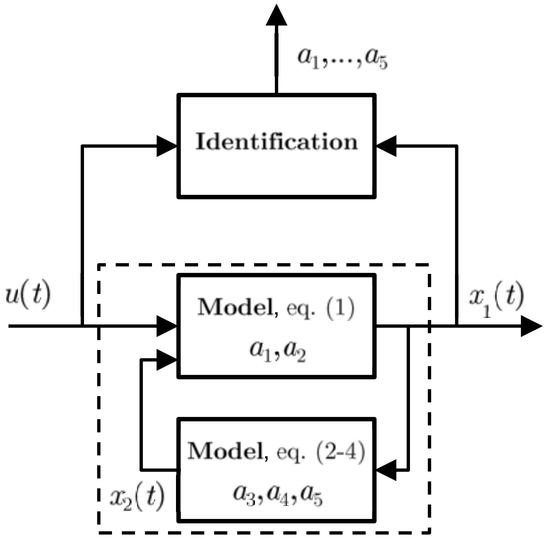


Fig. 4. System identification

Denoting $\mathbf{x} = [x_1 \quad x_2]^T$ and $\mathbf{a} = [a_1 \ \ldots \ a_5]^T$, we may write the model equations (1)–(3) in a compact form:

$$\mathbf{x}(t) = \Phi\left(\mathbf{x}(t), \mathbf{a}; u(t)\right). \tag{4}$$

Model Based Optimization of the Training Protocol (Planning). The role of proposed eHealth system is to manage, store, utilize and adjust mathematical models of exercisers. With use of the model adjusted to the exerciser, optimal training protocols are generated and actuated.

For a **given**: – distance D to run, – the model of exerciser, – fatigue limit $x_2^{\max}$ of exerciser (after this upper bound is exceeded, exercisers terminates a race), the task is **to find** such a training protocol $u^*(t)$, for which a desired distance is completed in a shortest time T^*. If necessary, we may take into account additional requirements concerning upper bounds for the heart rate ($x_1^{\max}$), heart rate change ($\Delta x_1^{\max}$) and lower bound for blood glucose level $G_{\min}$ during exercise.

Performance index Q is the function:

$$T = Q\left(u(t), D\right) \tag{5}$$

and it is solution of the equation:

$$D = \int_0^T u(t)dt \tag{6}$$

with respect T, for given D and $u(t)$. Fatigue constraint has the form:

$$\max_{0 \le t \le T} x_2(t) \le x_2^{\max}, \tag{7}$$

where x_2 is related to $u(t)$ by the model equation (4). If we want to force the exerciser to do his best, we may require, that the highest fatigue occurs at least at the end of the race:

$$x_2(T) = x_2^{\max}. \tag{8}$$

Moreover, the exerciser can not run faster than $u_{\max}$, due to the fitness level:

$$\max_{0 \leq t \leq T} u(t) \leq u_{\max}. \tag{9}$$

Additional requirements for people having cardiovascular problems may be included:

$$\max_{0 \leq t \leq T} x_1(t) \leq x_1^{\max}, \tag{10}$$

$$\max_{0 \leq t \leq T} \frac{d}{dt} x_1(t) \leq \varDelta x_1^{\max}, \tag{11}$$

and for diabetic:

$$\max_{0 \leq t \leq T} G(t) \leq G_{\min}, \tag{12}$$

where $G(t)$ is predicted track of blood glucose level, worked out from the model described in the work [5].

To sum everything up, for the basic setting (involving constraints (7)–(9), optimal training protocol $u^*(t)$ is solution of the following **optimization task**:

$$u^*(t) = \arg \min_{u(t) \in \mathscr{U}} Q\left(u(t), D\right), \tag{13}$$

where $\mathscr{U}$ is the space of all possible function $u(t)$ and the shortest time for the training protocol $u^*(t)$ is:

$$T^* = Q\left(u^*(t), D\right). \tag{14}$$

Inequality constraints $\boldsymbol{\psi}(u(t), \varPhi) \leq 0$ in the vector notation are:

$$\max_{0 \leq t \leq T} \begin{bmatrix} x_2(t) \\ u(t) \end{bmatrix} \leq \begin{bmatrix} x_2^{\max} \\ u_{\max} \end{bmatrix}, \tag{15}$$

where $x_2(t)$ is derived from the model $\varPhi$ (equation (4) and there is one equality constraint $g(u(t), \varPhi) = 0$:

$$x_2(T) = x_2^{\max}. \tag{16}$$

As stated before, the space $\mathscr{U}$ of all possible functions $u(t)$ is limited to composition of step-like functions, as shown in Fig. 4. The process of training protocol design may be simplified by parametrization of the function $u(t)$. Parameters should define: length of the periods (resting, exercise and recovery) and associated speeds (recovery period has zero speed by default). The order of periods determine their start and stop time instants. It is also possible to assign a predefined speed to each period, which stems from the fact, that people walk and run with their characteristic speeds. Parametrization makes optimization easier, but introduces the problem concerning the total number of parameters describing the solution $u(t)$. Since the footrace is completed after the distance D is done, the number of parameters may vary, depending on values of parameters. Thus, only optimization methods that may perform search the space with varying number of dimensions, such as Simulated Annealing or Evolutionary Algorithms, may be applied.

Training Support by the Control Algorithm. The training protocol $u^*(t)$ obtained by the optimization algorithm must be actuated. The exerciser is expected to follow this protocol. We have to support the exerciser in switching between exercise, resting and recovery periods in proper time and in maintaining the correct speed during these periods. This may be applied using voice commands uttered by the smartphone (e. g. 'run faster', 'slow down please') or by modifying music volume (if the exerciser likes to listen to the music). In the former case, we have on/off controller and in the second case there are many control actions that may be applied (but the space of control actions is still discrete).

The simplest control system structure possible is depicted in Fig. 5.

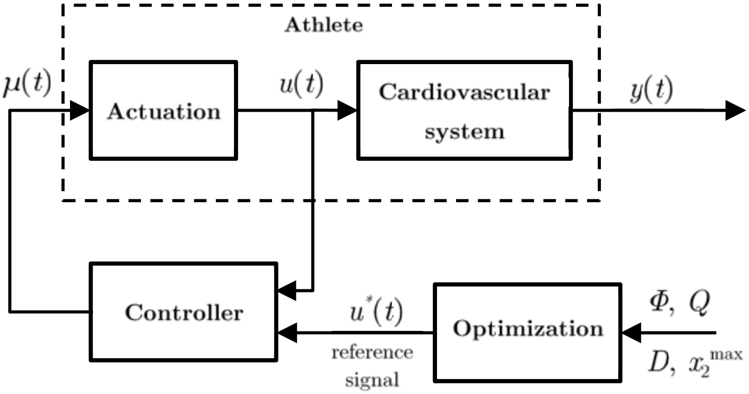


Fig. 5. Control system

Optimization algorithm takes the model Φ, performance index Q, parameters D and $x_2^{\max}$ and produces the best training protocol $u^*(t)$. The training protocol is accomplished by the exerciser with the aid of feedback controller (e. g. PID controller), which generates signals $\mu(t)$ (e. g. voice commands) in order to maintain the proper speed. Cardiovascular system response $y(t)$ is measured by the pulsometer and it may be further used to satisfy another training goals and requirements (e. g. keeping the heart rate within safe interval). In such a case we would obtain a hierarchical control scheme. We may also react to differences between the measured heart rate profile $y^*(t)$ and the profile $y^*(t)$ predicted by the model Φ. If there are significant differences, this means that it is very likely that the exerciser will not manage the optimal training protocol. Then it would be better to redesign the protocol by some adaptation procedure.

5 Summary

In the work a eHealth system to support planning training protocol for exerciser are presented. In order to solve introduced problem the system for distributed computing environment based on PAN and BAN and remote server is proposed. Benefits of the proposed architecture which are discussed in details are context-awareness and concept of personalisation. As it was stressed they impact on quality of usability of the system and improve management of network and computational resources.

Introduced eHealth system is employed to acquire and deliver data and signals for e.g.: identification, optimisation and control/management. In this work problem of support management of dynamic exercise intensity for proposed system is formulated. To this end relationship between heart rate and exercise intensity is applied. At the end possible solution by use of PID controller is discussed.

Acknowledgments. This work was partially supported by the European Union from the European Regional Development Fund within the Innovative Economy Operational Programme project number POIG.01.01.02-00-045/09-00 "Future Internet Engineering".

References

1. Cheng, T.M., Savkin, A.V., Celler, B.G., Su, S.W., Wang, L.: Nonlinear Modeling and Control of Human Heart Rate Response During Exercise With Various Work Load Intensities. IEEE Trans. On Biomedical Engineering 55(11), 2499–2508 (2008)
2. Cheng, T.M., Savkin, A.V., Celler, B.G., Su, S.W., Wang, L.: Heart Rate Regulation During Exercise with Various Loads: Identification and Nonlinear H_{inf} Control. In: Proc. of the 17th World Congress of The International Federation of Automatic Control, pp. 11618–11623. IFAC, Seoul (2008)
3. Dalla, M.C., et al.: GIM, Simulation Software of Meal Glucose-Insulin Model. Journal of Diabetes Science and Technology 1, 323–330 (2007)
4. Dalla, M.C., et al.: Meal Simulation Model of the Glucose-Insulin System. IEEE Transactions On Biomedical Engineering 54, 1740–1749 (2007)
5. Dalla, M.C., et al.: Physical Activity into the Meal Glucose-Insulin Model of Type 1 Diabetes: In Silico Studies. Journal of Diabetes Science and Technology 3, 56–67 (2009)
6. Grandinetti, L., Pisacane, O.: Web based prediction for diabetes treatment. Future Generation Computer Systems 27, 139–147 (2011)
7. Greene, B.R., McGrath, D., O'Neill, R., O'Donovan, K.J., Burns, A., Caulfield, B.: An adaptive gyroscope-based algorithm for temporal gait analysis. Journal of Medical And Biological Engineering And Computing 48, 1251–1260 (2010)
8. Johnson, D., Perkins, C., Arkko, J.: RFC: 3775. Mobility Support in IPv6. Technical report, Network Working Group (2004)
9. Lornicz, K., Chen, B., Challen, G.W.: Mercury: A Wearable Sensor Network Platform for High-Fidelity Motion Analysis (2009)

10. Mougiakakou, S.G., et al.: SMARTDIAB: A Communication and Information Technology Approach for the Intelligent Monitoring, Management and Follow-up of Type 1 Diabetes Patients. IEEE Transactions on Information Technology in Biomedicine 14, 622–633 (2010)
11. Twomey, N., Faul, S., Marnane, W.P.: Comparison of accelerometer-based energy expenditure estimation algorithms. In: 4th International Conference on Pervasive Computing Technologies for Healthcare (PervasiveHealth), pp. 1–8. IEEE Press, Munich (2010)
12. BodyTel, `http://www.bodytel.com`
13. LifeScan, `http://www.lifescan.com`
14. MyGlucoHealth, `http://www.myglucohealth.net`
15. World Health Organization, `http://www.who.int`

Detection of Facial Features on Color Face Images

Hsueh-Wu Wang[1], Ying-Ming Wu[2], Yen-Ling Lu[3], and Ying-Tung Hsiao[1]

[1] Department of Electrical Engineering, Tamkang University,
25137 Taipei, Taiwan
[2] Department of Digital Technology Design, National Taipei University of Education,
10671 Taipei, Taiwan
[3] Central Personnel Administration, Executive Yuan,
10051 Taipei, Taiwan
{hwwang,ythsiao}@tea.ntue.edu.tw
wuym@pchome.com.tw
anny@cpa.gov.tw

Abstract. This paper proposes a solution algorithm to locate the facial features on the human face images. First, the proposed algorithm determines the face region based on skin-tone segmentation and morphological operations. Then, we locate the facial features (i.e. brows, eyes, and mouth) by their color information. Finally, this algorithm set the shape control points based on the Facial Animation Parameters in the MPEG-4 standard on the located facial features. Results of experiments to the face images show that the proposed approach is not only robust but also quit efficient.

Keywords: Eye detection, face detection, face recognition, image processing.

1 Introduction

Detection of facial features is a major concern in a wide variety of applications such as human computer interaction, facial animation, facial expression, face recognition, and face image database management. Facial features extraction is nothing but identifying exact location of different features on face which include detection of eyes, brows and mouth etc. Among the various facial features, eyes, brows and mouth are the most prominent features in many applications. Specifically, eyes play an important role for the analysis on facial images. Hence, it is advantageous to locate eyes before the location of other facial features for detecting the facial features. After detecting eyes, the location of other facial features can be estimated by the eyes location.

Ahn and Ozawa developed a method based on estimation of muscular contraction parameters from facial feature points for generating facial expressions [1]. Zhang et al. located the facial features by defining subareas for facial features and comparing the geometry between these subareas for facial expression synthesis [2]. Song et al. proposed the skin deformation parameters based on the color changes regarding the corresponding area of the facial features [3]. [4-6] proposed the appearance methods

J.-S. Pan, S.-M. Chen, N.T. Nguyen (Eds.): ACIIDS 2012, Part I, LNAI 7196, pp. 86–101, 2012.

to detect facial features based on their photometric appearance [7-9]. Recently, more efficient methods have been introduced, e.g., the voting method by Lowe [10] and the statistical method by Weber [11]. Lowe has his own method for selecting and representing keypoints [10], but the applications of Weber's approach in [12] utilize unsupervised descriptors by Kadir [13].

This study utilizes color human face images for locating the facial features. Unlike other methods utilizing grey face images, color images contain more messages regarding the facial features to enhance the locating features. Color images are also good for various applications with improved result [13-17]. Therefore, to remove the effect coming from background or other objects of the non-face area, the first step is to divide the image into skin color and non-skin color areas so that the interesting areas of the face image can be cut from the imaged. This step can effectively reduce the computer burden and required follow-up of pixels. Also, the range of locating features can be limited in the area on the known color regions to improve the solution accuracy.

This paper presents a locating algorithm to determine the positions of the facial features on the human face images. First, the proposed algorithm determines the face region based on skin-tone segmentation and morphological operations. Then, we locate the positions of brows, eyes, and mouth by their color and geometry information. Finally, this algorithm makes the control points by the Facial Animation Parameters in the MPEG-4 standard on the located facial features.

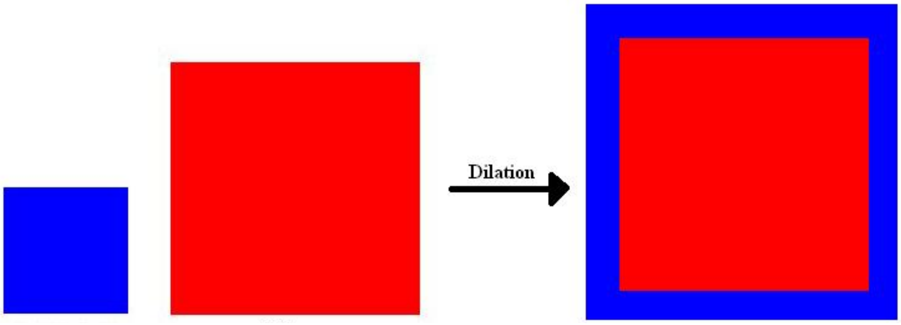
The remainder of the paper is organized as follows. In section II, we briefly review the method of feature-based object detection and localization. Section III presents the proposed solution algorithm to locating the facial features. Section IV devotes to the experiments on the face database. The last section gives the conclusion.

2 Proposed Method

This study utilizes CbCr color space to build the criteria for segmenting skin colors. YCbCr is a family of color spaces used as a part of the color image pipeline in video and digital photography systems. Y is the luminance component and Cb and Cr are the blue-difference and red-difference Chroma components. YCbCr commonly used in television produced film, video, image compression, different image signal processing between devices. Sometimes known as YPbPr, is a non-absolute color space, which the three components that cannot be accurate to form a color, so the need to convert the RGB color components of YCbCr space to the corresponding value. According to the literature [18] and [16], there is not with direct relationship between the skin of distribution and brightness in YCbCr color space. If the brightness is considered as the index for color segmentation, the non-uniform brightness distribution may affect the results on verdict. Therefore, the luminance component can be removed, not included in the calculation. The detection on different ethnic skin color in CbCr space at the same time within similar color can also be [17], with considerable robustness.

After the color image segmentation, the background due to other objects or light effects, there will be left a little under discontinuous and scattered blocks of varying size ratio on the images. If not dealt with these noises may conduct errors on the process images and also will increase the computation complexity. Therefore, this step uses morphological image processing operation including dilation, erosion, closing, area fill, the regional clean, bridge, Spur and Majority computing [15] to remove irrelevant, redundant blocks and the gap on connection, to fill the outline of the slender gap and the missing link finer pixels. Morphology with swelling, erosion of basic operations are built and used interchangeably for the follow-up processes. The implementation of these operations needs elements of a rectangular structuring element for processing on the images. These operation are divided into binary and grays images computing. As the operations on the two kinds of images with different definitions, the following definitions of the operations are dedicated to the binary image with detailed description where Z is defined as two-dimensional integer space, A is the object image for processing, and B is the structural elements.

The expansion opearation is defined as: A and B are two sets in the Z space, the B of A to B every time one direction along the center of the boundary moves along the perimeter of A, every move in all its original point for reflection and shift reflecting this z units, making the reflection of B and A overlap at least one element of the collection for utilizing the displacement of the B to A to expand the boundary. The shape and size of the structure element B can be defined as a rectangular or round and so on. Usually it is set in a 3 × 3 array as the smallest unit size with fill in the value 1 for the direction of operation, the rest as 0. A swelling expressed by B as in equation (3-4) and (3-5). Assuming the structure element B with length and wide as 2d unit, A's wide and length as 2x unit, the expansion is executed after every point along the circumference of A against B, extension center, making the A's side into 2 (x + d) units, as illustrated in Fig. 3-1.



structure element B image A results of the expansion

Fig. 1. The operation of expansion in Morphology

$$A \oplus B = \left\{ z \middle| \left(\hat{B}\right)_z \cap A \neq \varnothing \right\} \tag{1}$$

$$A \oplus B = \left\{ z \middle| \left[\left(\hat{B}\right)_z \cap A \right] \subseteq A \right\} \tag{2}$$

This experiment of this project utilizing the images with neutral expression instead of expression images because the positioning accuracy of the locating features will be affected by wrinkle on faces and the samples regarding expression images are not readily available. After segmenting face region image, the easily recognizable face for the eyes, eyebrows, mouth are set for the target features. These features was cut out and marked on their four corners of the right, left, upper and lower boundary points for the target. The choices of these boundary points are defined by reference to the MPEG-4 FAP feature points [19] and Song et al. [20] for locating the main affecting of the expression features as the action points.

According to the literature [15, 21, 22], from the point of view of gray scale images of human faces, we can found that the gray scale value in the pupil and iris is usually lower than that of the surrounding skin color, i.e. the color in the gray-scale image would be more black. With this feature, the RGB color image (that the region has been previously cut) was converted to gray-scale image. By setting a threshold value, the gray-scale image was mapped into binary image for separating the eye area out from the image block. This study used a single threshold value to transfer the image into binary image. We set the initial value of the threshold value as T_0, maximum T_{max}, and a fixed increasing value T_{step}. Between T_0 to T_{max} with the different thresholds to images for transferring into binary pattern, each value of pixels larger than the threshold value is set to white, the rest of the pixels set to black. There are $(T_{max} - T_0) / T_{step}$ sheets of image produced in this step. Then the establishment of rules is applied to analyze each image in the black area (i.e. below the threshold). By these rules, this process eliminates unnecessary blocks to determine which of the two regions may be the position of the eye. Its rules by two geometric position relationship building as follows:

(1) The horizontal distance between the centers of the two eyes will be limited in a fixed-pixel range.
(2) The vertical distance between the centers of two eyes will be limited within a pixel range.
(3) The block size of two eyes with its length and width will be limited to a ratio of pixel range.

After filter out the eyes of the candidate blocks, this algorithm use 2-D correlation coefficient to calculate the degree of similarity between blocks. Before calculating correlation coefficient, because two of the shape is symmetric, one of the blocks should be flipped for relatively high similarity. The 2-D correlation coefficient represents the similarity of two blocks, coefficient does not have a specific unit, and its value in the range of +1 to -1. Its value closer to 1, the relationship that the more similar for the two variables; otherwise closer to 0, indicating that two variables are less similar. The positive value of a 2-D correlation coefficient imply positive change of a variable with the other variables will also increase at the same time, i.e. a positive relationship between the two variables. The 2-D correlation coefficient with negative value represents an increase in one variable while the other variable will be reduced, in other word the relationship with reverse. It is noteworthy that the closer to 1 the value of a positive sign.

The following rules are utilizing to locate the eye area:

(1) If r is greater than or equal to 0.5, the corresponding area can be identified as the eyes.
(2) If r is less than 0.5 but greater than 0, the corresponding area can be identified as the eyes.
(3) If r is less than or equal to 0, the corresponding area can be identified as the eyes.

According to the eye location coordinates (Eye_C_X_L, Eye_ C_Y_L; Eye_ C_X_R, Eye_ C_Y_R) determined by the above rules, the position and the color relationships between the eyes and eyebrows, the rules to find the position of the eyebrows are presented.

(1) The location of eyebrows must be above the eyes.
(2) The pixel of the vertical gap between the eyebrows and eyes is in a certain range.
(3) The width of the eyebrows is usually longer than the width of the eye within a certain pixel.

According to the above rules, we can create a mask to limit the search in the eyes of some of the top non-color region. The area identified as the location of the eyebrows, and the boundary points marked on the image, the positioning step is completed. This study does not directly use the characteristics of color to search eyebrow in order to avoid the light of their impact [15], which leads to search hard to control. Coordinates to get the eyebrows (Brow_X_L(i, j, k), Brow_Y_L(i, j, k); Brow_X_R(i, j, k), Brow_Y_R (i, j, k)) where the Brow_X is the eyebrows of X coordinates, Brow_Y is for the Y coordinates, subscript L, R, represent lift and right, respectively, i, j, k define the coordinates of boundary points.

During the experiment on the skin segment, the mouth region has its specific color distribution observed in the YCbCr color space, and surrounding skin are obvious differences, and took possession of a certain size of the pixel area. We apply this feature to the mouth region segmentation and mark their location boundary points. The results from the observation and found the mouth of the boundary points will occur the phenomenon of migration, and expected some drop. To reinforce positioning accuracy, the lip of the mouth, the cross point of the lips and the midpoint of this line are utilized to assist correct boundary points. The lip in the HSV color space is particularly evident and in the literature [21] also proved this view point. After segmenting the lip line according to the differences in color, the original boundary line is change to the midpoint of the calibration reference. So to get the right boundary point to (Mou_X (i, j, k, l), Mou_Y (i, j, k, l)) where Mou_X is the X coordinate of the mouth, Mou _Y is the Y-coordinates, i, j, k, l denote the coordinates of boundary points.

3 Experimental Results

This study utilizes Georgia Tech Face Database [23] for testing. The database contains fifty images of different people each with fifteen different angles, lighting, and expression of images. These images are the size of 640*480 JPEG color format. We select positive and expressionless images to test the proposed algorithm.

According to the process presented in section 3, after loading the image processing for color image segmentation, test images is studied with different color space histogram to obtain the best color distribution space. On color images in different color spaces and more layers of different shades of gray that the composition of the image histogram statistics of the different range of pixel gray levels, these information are get to understand the concentration and distribution trends. Figure 2, 3 and 4 illustrate the typical test examples from Georgia Tech Face Database.

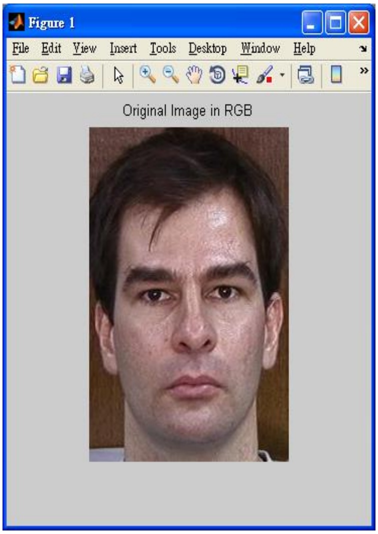


Fig. 2. Test pattern 1

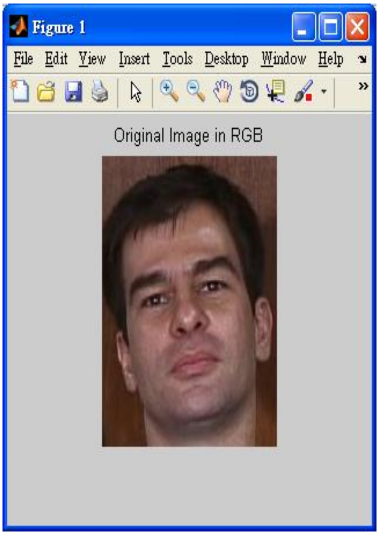


Fig. 3. Test pattern 2

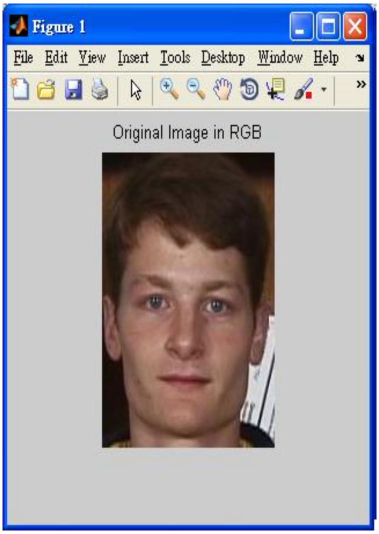


Fig. 4. Test pattern 3

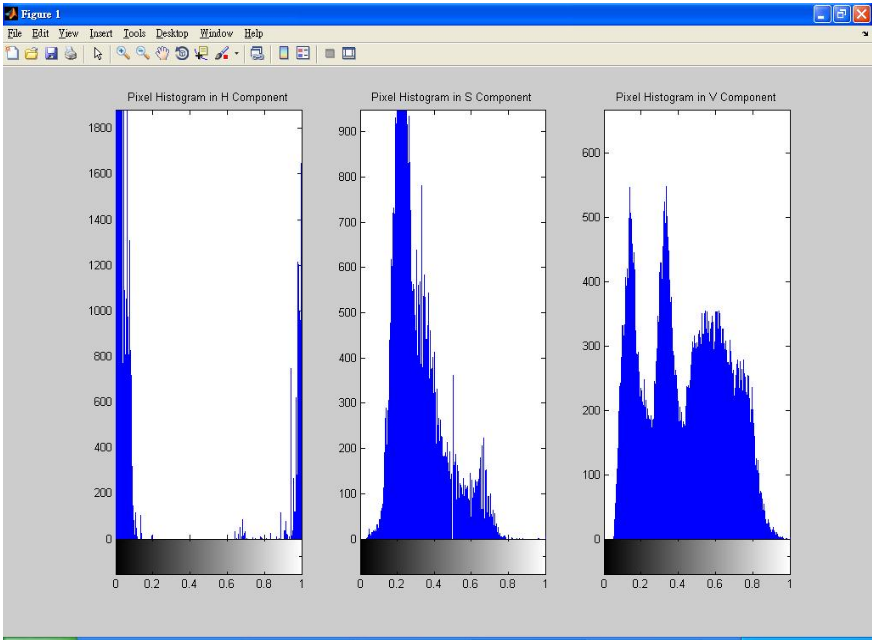


Fig. 5. HSV color space histogram statistics on Fig. 2

Fig. 5 shows the HSV color space histogram statistics on Fig. 2. From the Fig. 5, the distribution regarding the saturation component S concentrated in the 0.2 to 0.4 with a certain trend. The gray value distribution is more extreme due to the effect of color component H. Hence, the search will obviously cover the whole HSV space

[0, 1] making choice of unnecessary pixels. Figure 6 displays the original image conversion to HSV space and Figure 7 shows the segment of HSV color space.

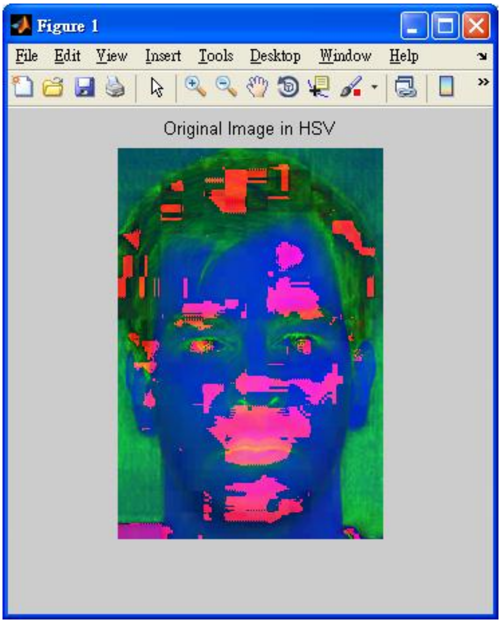


Fig. 6. The original image conversion to HSV space

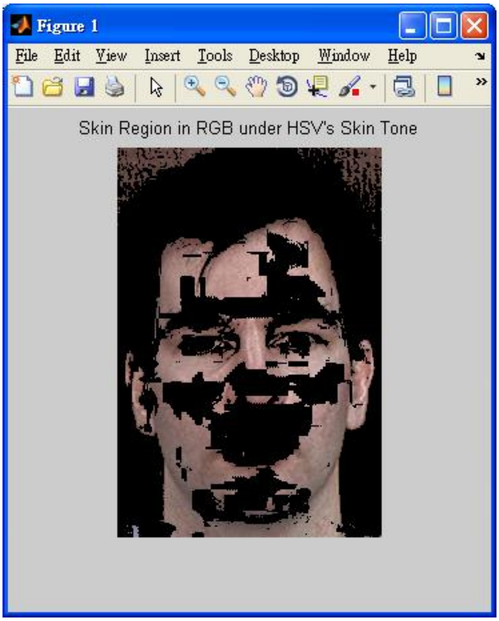


Figure 1
File Edit View Insert Tools Desktop Window Help
Skin Region in RGB under HSV's Skin Tone

Fig. 7. The segment of HSV color space

From Fig. 7, the color segmentation of images effect by light and shadow resulting in many sub-blocks belonging to the face area not to be identified as the color and also resulting in the presence of noise. In order to maintain the integrity of the face for locating features, it is need to filter noises from images by the morphology operations. The morphology operations are logical operating. In order to facilitate the process, the color image transfer into binary image shown in Fig. 8. Fig. 9 exhibits the filter results.

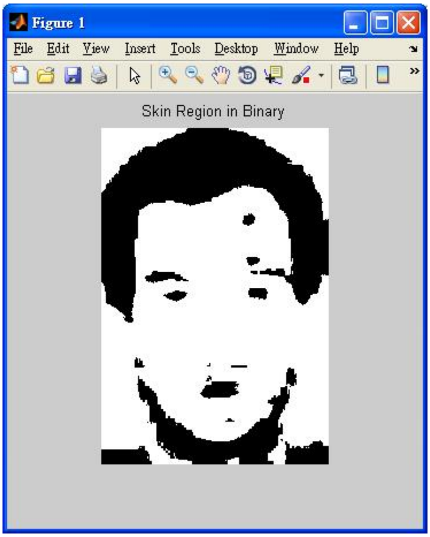


Fig. 8. The color image transfer into binary image

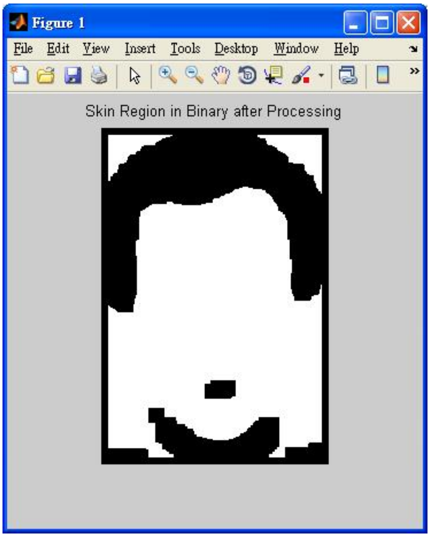


Figure 1
File Edit View Insert Tools Desktop Window Help
Skin Region in Binary after Processing

Fig. 9. The filter results of Fig. 8

The colors of pupil and its around area are with difference gray value. Therefore, we segment image through setting suitable threshold value. Refer the suggestion in [15], the threshold can set at the range between 0.1 to 0.6. From the experiments, while the threshold value is greater than 0.3, the segmentation image information in usability will gradually become less and has become more difficult for future process. The optimal threshold is set in the range of 0.13 to 0.22 for eye segmentation. For conservative estimate, the range of threshold is set in 0.1 to 0.3. The next step followed by the image segment is the establishment of rules to determine which blocks should be removed due to useless. Because the locations of eyes and face are inside the face, we can determine the black block covered with white block to be filtered out many of the extra block.

After determining which regions are inside the face, the rules of the geometric relationship between two eyes are developed to determine which block blocks are eyes as follows.

(1) The horizontal distance of center of two blocks within the level 45 to 75 pixels.
(2) The difference on the vertical distance of center of two blocks is less than 20 pixels.
(3) The range of two blocks is within 40 × 20 size.

Then calculating the correlation coefficient r of the two candidate blocks is for judging eyes. If r is greater than 0.5, the two blocks can be identified as the eyes. If r is less than 0 the two blocks will not be identified as the eyes. Moreover, If the value of r between 0.5 and 0, selecting the maximum value of r corresponding to the block identified as the eyes. Finally, the center will mark on the two blocks as shown in Fig. 10.

In fact, the locations of the eye points (detected by the above rules) are quite near. There are only differences in the decimal point. Therefore, the results of all the coordinates are obtained by rounding to integer. The coordinates with the highest frequency outcome is set as the eye position shown in Figure 11.

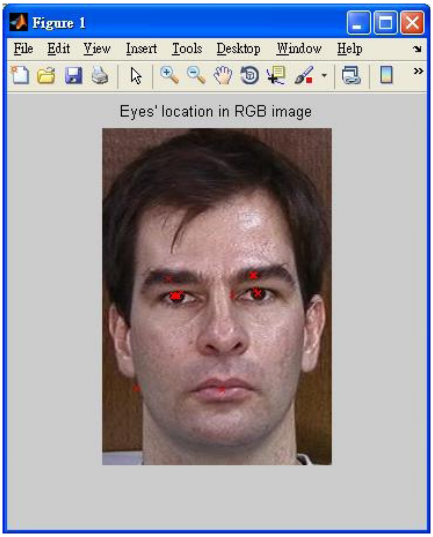


Figure 1
File Edit View Insert Tools Desktop Window Help
Eyes' location in RGB image

Fig. 10. The center of the located tagets marked on the blocks

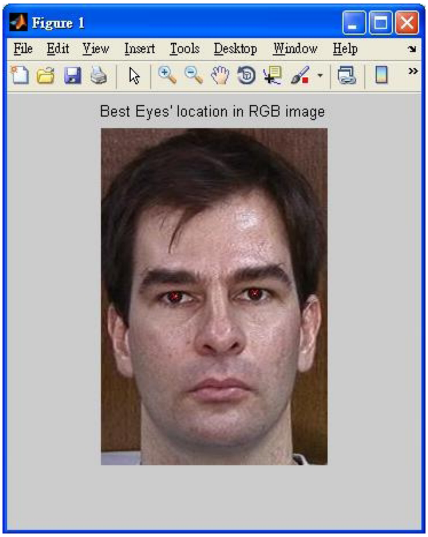


Figure 1
File Edit View Insert Tools Desktop Window Help
Best Eyes' location in RGB image

Fig. 11. The located results of eyes

The following rules describe the positioning of the eyebrows are utilized to locate the eyebrows' position.

(1) Eyes should be above the eyebrows.
(2) The vertical gap within between the eyebrows and eyes is about 20 pixel.
(3) The width of eyebrows usually is higher than the width of the eye within approximately 30 pixels.

The above rules (1) limite the searching in the area above eyes only, (2) only searching the region with non-color pixel value, and (3) stores the result by the binary image format. with the formation of a mapping diagram. Figure 12 shows the position of the eyebrows represented as white non-color region. Figure 13 is the located position marked with white color.

Final disposal of the locating feature is the positioning of the mouth, the same use in the YCbCr space around the mouth and skin color of the color differences, to segment the mouth differently. However, due to use of the Y component, it is easy to select the non-mouth area to the rest of the region of mouth shown in Fig. 14. Therefore, morphological operations required to dispose of the extra little noises to get the full outline of the mouth area as shown in Fig. 15. The following rules are to locate mouth on the binary image of white connected set with noise treatment.

(1) The position of mouth is at 1/2 to 3/4 of the face image.
(2) The area of mouth is usually above 900 pixels in the face image.
(3) The ratio of ratio of width to length on mouth is around the range 2:1 to 3:1.

After further screening by the rules, the final determination of the mouth area shown in Fig. 16. Although its shape is not the complete line shape of the mouth, somewhat prominent and distorted edges are generated, the outline is in the acceptable range. Finally, the results of the located position of mouth map to the RGB space, the images show the locating region of the mouth as in Fig. 17.

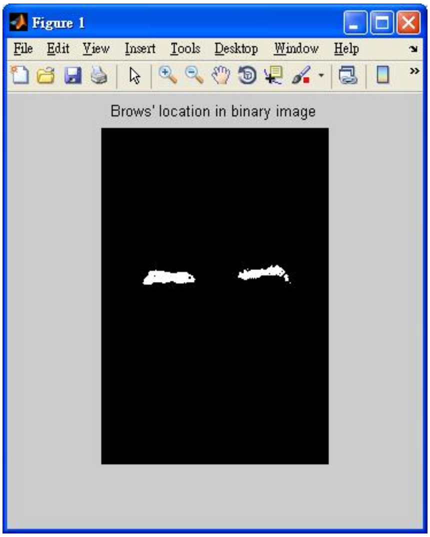


Figure 1
File Edit View Insert Tools Desktop Window Help
Brows' location in binary image

Fig. 12. The position of the eyebrows represented as white non-color region

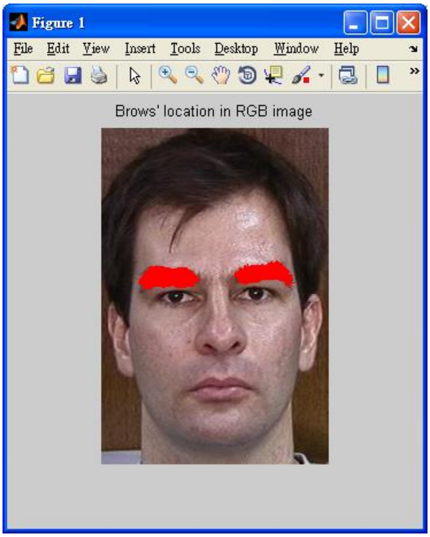


Figure 1
File Edit View Insert Tools Desktop Window Help
Brows' location in RGB image

Fig. 13. The positions of the located eyebrows

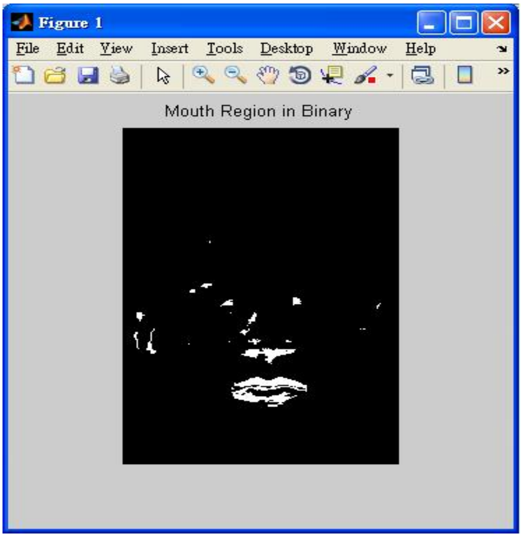


Fig. 14. The segment mouth of facial image

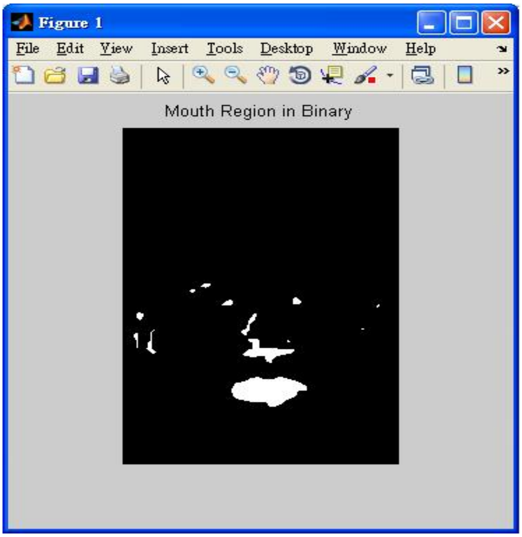


Fig. 15. The filter noise of Fig. 4-27

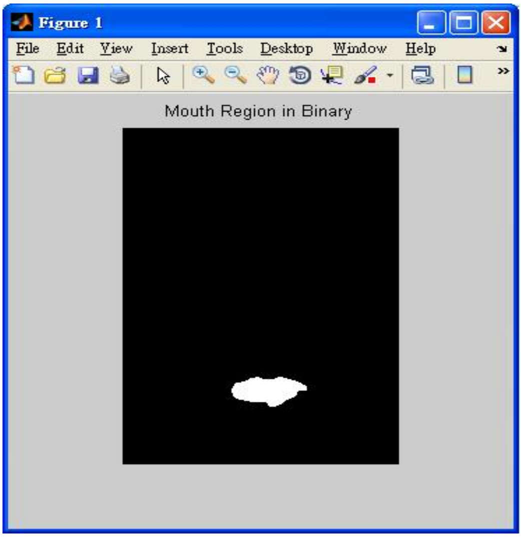


Fig. 16. The located mouth on the binary image of white connected set with noise treatment

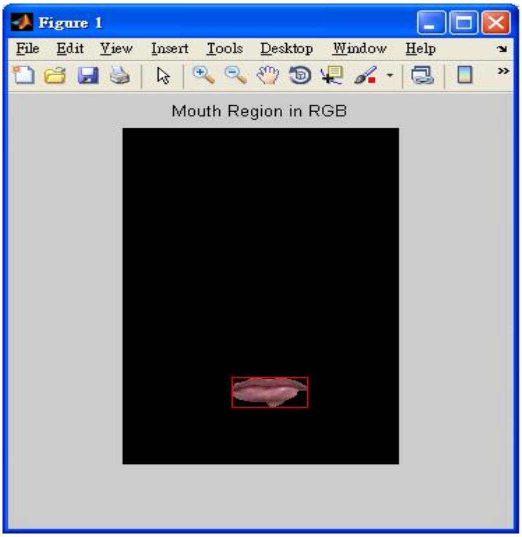


Fig. 17. The located region of the mouth

4 Conclusions

This paper proposed a face detection algorithm for facial color images using a skin-tone color model and facial features. First, the proposed algorithm determines the face region based on skin-tone segmentation and morphological operations. Second, we locate the facial features by their color and geometry information. Finally, this algorithm set the shape control points according to the Facial Animation Parameters in the MPEG-4 standard on the located facial features. The proposed method has been tested by images from Georgia Tech Face Database. Experiment results show that this method works well with the facial images.

References

1. Ahn, S., Ozawa, S.: Generating Facial Expressions Based on Estimation of Muscular Contraction Parameters From Facial Feature Points. In: IEEE International Conf. on Systems, Man, and Cybernetics, The Hague, Netherlands, October 10-13, vol. 1, pp. 660–665 (2004)
2. Zhang, Q., Liu, Z., Guo, B., Terzopoulos, D., Shum, H.Y.: Geometry-Driven Photorealistic Facial Expression Synthesis. IEEE Transactions on Visualization and Computer Graphics 12(1), 48–60 (2006)
3. Song, M., Tao, D., Liu, Z., Li, X., Zhou, M.: Image Ratio Features for Facial Expression Recognition Application. IEEE Transactions on Systems, Man, and Cybernetics – Part B: Cybernetics 40(3), 779–788 (2010)
4. Pentland, A., Moghaddam, B., Starner, T.: View-based and modular eigenspaces for face recognition. In: Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR 1994), Seattle, WA (1994)
5. Min Huang, W., Mariani, R.: Face detection and precise eyes location. In: Proc. Int. Conf. on Pattern Recognition, ICPR 2000 (2000)
6. Huang, J., Wechsler, H.: Eye detection using optimal wavelet packets and radial basis functions (rbfs). Int. J. Pattern Recognit. Artif. Intell. 13(7), 1009–1025 (1999)
7. Chellappa, R., Wilson, C.L., Sirohey, S.: Human and Machine Recognition of Faces: A Survey. Proc. IEEE. 83, 705–740 (1995)
8. Lam, K.L., Yan, H.: Locating and Extracting the Eye in Human Face Images. Pattern Recognition 29(5), 771–779 (1996)
9. Smeraldi, F., Carmona, O., Bigun, J.: Saccadic Search with Gabor Features Applied to Eye Detection and Real-time Head Tracking. Image and Vision Computing 18(4), 323–329 (2000)
10. Lowe, D.G.: Distinctive image features from scale-invariant keypoints. Int. J. Comput. Vis. 60(2), 91–110 (2004)
11. Weber, M.: Unsupervised learning of models for object recognition. Ph.D. dissertation, California Inst. Technol., Pasadena (2000)
12. Fergus, R., Perona, P., Zisserman, A.: Object class recognition by unsupervised scale-invariant learning. In: Proc. IEEE Computer Society Conf. Computer Vision and Pattern Recognition, pp. 264–271 (2003)
13. Kadir, T.: Scale, saliency and scene description. Ph.D. dissertation. Oxford Univ., Oxford, U.K (2002)
14. Phung, S.L., Bouzerdoum, A., Chai, D.: Skin Segmentation Using Color Pixel Classification: Analysis and Comparison. IEEE Transactions on Pattern Analysis and Machine Intelligence 27(1), 148–154 (2005)
15. Tao, L., Wang, H.B.: Detecting and Locating Human Eyes in Face Images Based on Progressive Thresholding. In: Proc. of the 2007 IEEE International Conf. on Robotics and Biomimetics, Sanya, China, December 15-18, pp. 445–449 (2007)
16. Berbar, M.A., Kelash, H.M., Kandeel, A.A.: Faces and Facial Features Detection in Color Images. In: Proc. of the Geometric Modeling and Imaging –New Trends, July 05-06, pp. 209–214 (2006)

17. Guan, Y.: Robust Eye Detection from Facial Image based on Multi-cue Facial Information. In: Proc. of the 2007 IEEE International Conf. on Control and Automation, Guangzhou, China, May 30- June 1, pp. 1775–1778 (2007)
18. Gonzalez, R.C., Woods, R.E.: Digital Image Processing, 2nd edn. Prentice Hall, New Jersey (2002)
19. ISO/IEC Standard 14496-2, Coding of Audio-Visual Objects: Visual (October 1998)
20. Song, M., Tao, D., Liu, Z., Li, X., Zhou, M.: Image Ratio Features for Facial Expression Recognition Application. IEEE Transactions on Systems, Man, and Cybernetics – Part B: Cybernetics 40(3), 779–788 (2010)
21. Ding, L., Martinez, A.M.: Features versus Context: An Approach for Precise and Detailed Detection and Delineation of Faces and Facial Features. IEEE Transactions on Pattern Analysis and Machine Intelligence 32(11), 2022–2038 (2010)
22. Vezhnevets, V., Sazonov, V., Andreeva, A.: A Survey on Pixel-Based Skin Color Detection Techniques. In: Proc. of Graphicon 2003, Moscow, Russia, pp. 85–92 (September 2003)
23. Georgia Tech Face Database, http://www.anefian.com/face_reco.htm

Adaptive Learning Diagnosis Mechanisms for E-Learning

YuLung Wu

Department of Digital Content and Technology,
National Taichung University of Education
140 Min-Shen Road, TaiChung 40306 Taiwan R.O.C
ylw@mail.ylw.idv.tw

Abstract. In class teaching with a large number of students, teachers lack sufficient time in understanding individual student learning situation. The framework of learning activity in this study is based on the Learning Diagnosis Diagram. Before conducting learning activities, teachers must prepare Learning Diagnosis Diagrams. This work proposes an adaptive Learning Diagnosis Diagram to consider differences among students. The proposed system provides a personalized Learning Diagnosis Diagram for individual students and adjusts learning phases to automatically fit student achievement. The learning evaluation demonstrates the effectiveness of the proposed method. The evaluation shows that the Learning Diagnosis Diagram can provide an adaptive learning environment for students.

Keywords: Learning Diagnosis, e-Learning.

1 Introduction

In on-line distance learning environments, Internet is the major media for communication between instructors and learners. The methods most frequently used by instructors to understand the situation of learners are setting homework or examinations. However, these methods do not provide instructors with immediate feedback, and the feedback obtained inevitably lags behind teaching progress. The lack of face-to-face interaction between instructors and learners in on-line distance learning makes it difficult for instructors to know the situation of learners, and this weakness must be improved. By recording and analyzing learner behavior, details of individual learning situation can be obtained, including progress, strengths and weaknesses. Furthermore, a system can be designed to automatically generate a series of aided-learning contents for every student.

In Taiwan, in classroom environments with a large number of students, teachers rarely have time to monitor student status. As a result, students who learn quickly thoroughly grasp the content, while students who learn slowly fall further and further behind until, sometimes, the education system gives up on them.

This research proposes a novel learning process for distance learning environments. The novel process is based on the Learning Diagnosis Diagram (LDD), a knowledge representation tool. The LDD contains learning content and guides the learning activities. The LDD changes its structure automatically to suit the conditions

J.-S. Pan, S.-M. Chen, N.T. Nguyen (Eds.): ACIIDS 2012, Part I, LNAI 7196, pp. 102–110, 2012.

of each student. A good student only needs to learn less but more important learning content. Conversely, a poor student must study content in detail and evaluate their weaknesses. Teachers need only develop full and detailed course content. The proposed algorithm handles the different styles of students automatically in the course.

2 Related Works

The "Conceptual graph" [8][9][10] is a traditional assisted teaching method. A conceptual graph is composed of propositions defined by two concept nodes and one connecting relation link. The concept nodes are arranged in a hierarchical structure. The conceptual graph can effectively represent knowledge structure. The common method applied in the education field is “Concept Mapping”, which allows learners to draw their conceptual graphs. Diagnoses are made by comparing the concept maps of each student with those of the instructor. The comparison results can be used to determine student comprehension levels for specific knowledge themes. Novak[8] proposed “Concept mapping” to score the comparison between instructor and each student. Goldsmith[2][3] improved the method and proposed “Pathfinder scaling”. Ruiz-Primo[9] evaluated two mapping methods: fill-in-the-map and construct-a-map. Construct-a-map scores reflected the differences among knowledge structure of students.

Hwang[4] designed a computer-assisted diagnostic system that includes concept-effect relationships. The concept-effect relationship is a set of association levels (Test Item relationship table, TIRT) that represents the relationship between each concept node and each test item. In Hwang’s research, the instructor must set an association level for each relationship before the course begins. According to the answers of students, the algorithm could obtain student learning information with each concept node. The existence of m concepts and n test items is assumed. The instructor then must focus on m×n association levels. Generally, a learning topic includes approximately 10~20 concepts and more than 200 test items.

Chang, Hung, & Shih[1] also proposed a similar approach for diagnosing student performance via a courseware diagram. The courseware diagram contains course units and evaluation units. A course unit represents a teaching process and an evaluation unit represents an evaluation process for a learning activity. This research designed different levels of difficulty for course units for a topic, including normal course, basic remedial course and enhanced remedial course. According to evaluation results, the evaluation unit chooses the suitable course unit as the next learning unit to continue the learning activities.

This research develops a novel strategy embedded in a computer-assisted system for diagnosing student learning situations and provides an adaptive learning environment. This research designed an on-line learning system with the proposed strategy. After students login to the learning server via the Internet, the system constantly monitors student behavior to produce learning portfolios for storage in the database server. The learning portfolios of each student include learning duration, login frequency, Q&A, discussion content, homework and statistical data. These learning portfolios enable the finding of students' learning achievement, and enable the automatic building of adaptive learning materials.

3 Learning Diagnosis Diagram

In this research, the learning activities of all students are guided by the proposed system. This research adopts Learning Diagnosis Diagram(LDD) as the basis of the proposed algorithm. First of all, the main definition is given.

Definition 1. An LDD D (C, R) is a direct and finite connected graph, where C denotes a set of nodes that indicates student study activities and R is a set of relationships. A relationship, r_{ij} , connects two distinct nodes c_i and c_j, where $c_i, c_j \in C$. The relationship represents the studying order, that is, the node c_i is learned before node c_j. Thus c_i is the prior-knowledge of c_j. In the algorithm, $r_{ij} = \{\overrightarrow{c_i c_j} \mid \forall c_i, c_j \in C, i \neq j\}$.

The LDD comprises teaching materials and test, where the LDD is defined by experts or teachers, and each learning node in the LDD has a related teaching material and test that are designed by teachers. From Fig. 1, a link exists between learning node 1 and 4, and the direction of the arrow represents the learning sequence. In the definition presented here, learning node 1 is a prior-knowledge of node 4. Similarly, students should learn learning node 1 before node 4 in the learning activities.

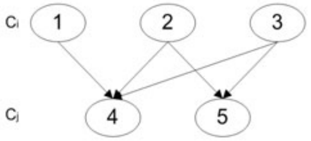


Fig. 1. LDD maps to the knowledge theme

The LDD is created by the teacher before the start of the course. The architecture of the LDD follows the course plan, including course contents and course sequence. This research adopts the course “Electronic Circuits Laboratory” as the evaluation course. The course is taught at the sophomore level. The "Electronic Circuits Laboratory" course introduced electronic devices and instruments used in circuits, and provided students with opportunities to implement circuit theories. The course familiarized students with basic theories and virtual circuits. The detailed information of the evaluation is described in section 4. Fig. 2 shows the LDD in this evaluation. The LDD was created by the course teacher.

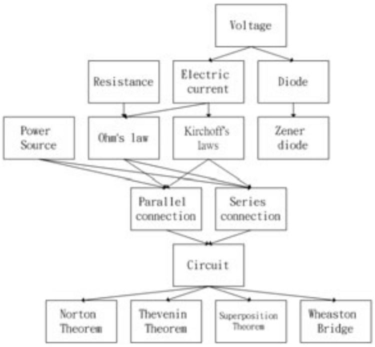


Fig. 2. LDD in the "Electronic Circuit Laboratory" course

In our recent research [5][6][7], LDD was developed and obtained positive results in three learning evaluations with amount 197 participants. To analyze response from teachers, one of important conclusion is the design of LDD is only considering many students, not all students. In other words, it is difficult to fit all students for one LDD. Before any learning activities, the teacher must prepare a LDD. During learning activities, students are guided by the LDD. The diagram includes all course content and course calendar. But in a large class, there are many students with different characteristics. Students with high-level ability learn material quickly and systematically. Students who belong to this category do not need to learn detailed information. A more effective method for these students is learning more about important topics. Conversely, poor students learn step by step and acquire the necessary information about a topic and, then, check their progress.

Using the collecting data and learning portfolio, the LDD records the performance for each student and is used to analyze the characteristics of their learning status. The main notion of the diagram is to identify similar learning content and combine them into one learning node. According to Fig. 2, the LDD is an initial diagram designed by instructors. The initial diagram contains the maximum number of learning nodes, implying the maximum number of learning steps. Students with a poor performance must learn diligently. These students learn and follow the LDD sequentially. However, detailed learning steps are not always advantageous for high performing students. For students with varying learning performances, learning plans with appropriate styles must be developed.

Prior to learning activities, students are divided into groups based on their learning performance of prior-test. Good students are guided by the LDD with fewer learning nodes. Poor students are guided by the LDD with more learning nodes. Students then begin their learning activities. To adapt to the individual needs of each student, the LDD adjusts automatically with the following algorithm. A student that fails on a learning node implies difficulty of the learning node for the student. Further information must be obtained to identify the cause of failed learning nodes. The learning node that the student fails, which is combined with similar learning nodes, is expanded. The student must thus complete the expanded nodes to resolve the exact learning problem.

The following is the definition of algorithm that used in the LDD.

Definition 2. Item Success index (ISI) represents the success learning of one learning node. To measure the degree of success learning, this research considers the ratio of right answers to total answers for one learning node. Value $ISI_{A,i}$ represents the item success index of learning node A for student i and ISI_A is the average success index of learning node A for all students. $ISI_{A,i}$ = (The number of right answers of student i / the number of answers of student i) * 100%, and $0 \leqq ISI_{A,i} \leqq 1$. $ISI_A = \frac{1}{n}\sum_{i=1}^{n} ISI_A$ when there are n students in a class.

An easy learning node often means many students can learn this learning node successfully. This research collects all answer of each learning node and calculates the correct radio, ISI. A small ISI_A value denotes that learning node A fails and hard to learn, meaning the learning node is necessary. And a large ISI_A value denotes the learning node is easy, meaning learning node can be ignored: i.e., if the learning node is easy, that means the learning node is not critical to students who perform well. This research adopts ISI as a factor to combine the easy learning nodes into one learning node.

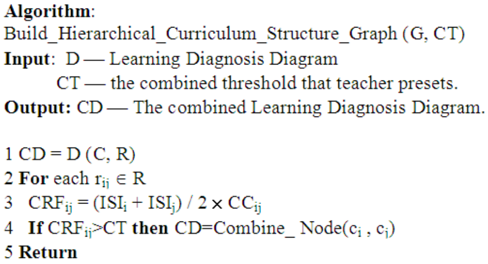
Definition 3. The correlation coefficient (CC) represents the correlation between two learning nodes. CC_{AB} represents the CC between learning nodes A and B. $CC_{AB} = \frac{\sum_{i=1}^{n}(ISI_{A,i} - ISI_A)(ISI_{B,i} - ISI_B)}{\sqrt{\sum_{i=1}^{n}(ISI_{A,i} - ISI_A)^2 \sum_{i=1}^{n}(ISI_{B,i} - ISI_B)^2}}$. Since negatives are not meaningful in this study, they are ignored in the calculations. Therefore, $0 \leqq CC_{AB} \leqq 1$.

The correlation value represents a degree of correlation, ranging from 1 to 0. The value of 1 denotes that the data of these groups are the same; a 0 value shows that these groups have no relationship with each others. From the answers, CC_{AB} shows the degree of similarity for learning nodes A and B. A larger CC_{AB} value denotes that students have the same learning performance, or we can say the attribute of two learning nodes is similar. That means two learning nodes contain similar or related content. This research assumes that two learning nodes that have a large CC_{AB} value are similar learning nodes. This research therefore combines similar learning nodes into one learning node.

Definition 4. Learning node Correlation Factor - The Learning node correlation factor (LCF) between learning node A and B, LCF_{AB} = ((ISI_A+ISI_B) / 2 × CC_{AB}, and $0 \leqq LCF \leqq 1$. The LCF_{AB} represents the degree of correlation for learning node A and B by evaluating success and similarity.

Definition 5. Combined Threshold – A combined threshold (CT) determines which learning nodes will be combined. If LCF_{AB} >CT, learning node A and B are combined into a single new learning node. Otherwise they are remained as separated.

This research utilizes CT to determine which learning nodes are combined. The CT value can be adjusted by teachers. A large CT contains more learning nodes, and a small CT contains fewer learning nodes. Teachers can preset some CTs that belong to students with different abilities. This proposed strategy generates automatically these related LDDs. The proposed algorithm is shown in Fig. 3.



```
Algorithm:
Build_Hierarchical_Curriculum_Structure_Graph (G, CT)
Input:  D — Learning Diagnosis Diagram
        CT — the combined threshold that teacher presets.
Output: CD — The combined Learning Diagnosis Diagram.

CD = D (C, R)
For each $r_{ij} \in R$
  $CRF_{ij} = (ISI_i + ISI_j) / 2 \times CC_{ij}$
  If $CRF_{ij}$>CT then CD=Combine_ Node($c_i$ , $c_j$)
Return
```

Fig. 3. The algorithm to build a LDD

The learning process designed in this study is shown in Fig. 4. Each learning node contains two steps. The first step is the learning step. In this step students must study the on-line learning materials. The second step is testing. During this step students take exams related to the learning materials. If students pass the test, they continue to study the next learning node according to their LDD. If students fail the test, it means the learning node is difficult, and thus students require easier and more detailed learning materials. The learning system expands the failed node that is combined with CT. Students who fail the test learn these expanded nodes to determine real problems in their learning activities.

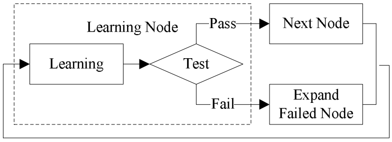


Fig. 4. Learning process of the LDD

With regard to learning activities, this study focuses on the "Testing step". The adaptive testing in the LDD is designed for students with different learning styles. The main idea of this research is detailed information for poor students and reduced information for good students. The detailed information indicates the expanded learning node while the reduced information indicates the combined learning node. The following is an important issue: "Does a test result from a combined learning node equal the test results from various expanded learning nodes?". The following section proposes a learning evaluation to discuss this issue.

4 Learning Evaluation

This section presents learning evaluation of this research. The participants were eighty students. All of the research participants had experience of using computers and the Internet. Each learning node has ten questions for test, so there are 150 questions in test item bank. All questions are multiple-choice questions.

The LDD in this research was constructed by two professors who had taught the course for many years, and is shown in Fig. 2. All learning nodes were taught in this evaluation in a semester. When designing the course learning activities, team assignments, conversation time and team examinations were arranged on a weekly basis. In the evaluation activities, all participants must participate in the instruction provided by the instructor. Meanwhile, in extracurricular activities, all participants login to the system and complete some review tests before taking the next course.

Figure 2 is a LDD with CT=1 and is a full LDD. The LDD contains all learning nodes and teaching materials and is suitable for low-level ability students. In Fig. 5a to 5c, the number of learning nodes was reduced progressively. Many similar learning nodes are combined progressively. Those learning nodes not combined were more important and difficult. Good students can learn by using the small LDD. From the results in Fig. 5a to 5c, the combining process also showed that those basic learning nodes (upper) and advanced theories and learning nodes (lower) of Electronic Circuits were retained. At the center of a diagram, there are basic theories which were combined together. In Fig. 5c, all learning nodes were grouped into three learning nodes: basic concepts, basic theories and advanced theories. The LDD generates different styles of learning content according to the ability of each student.

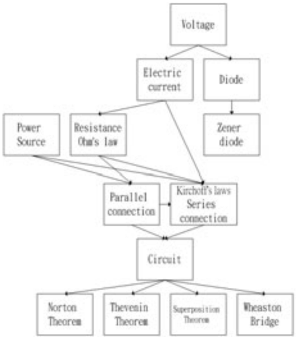


Fig. 5a. CT=0.55

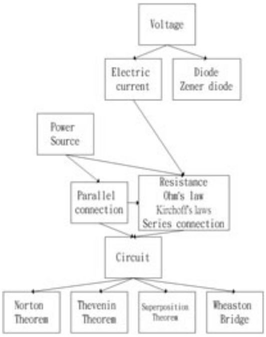


Fig. 5b. CT=0.5

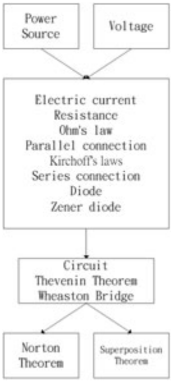

Power Source
Voltage
Electric current Resistance Ohm's law Parallel connection Kirchoff's laws Series connection Diode Zener diode
Circuit Thevenin Theorem Wheaston Bridge
Norton Theorem
Superposition Theorem

Fig. 5c. CT=0.45

This study discusses the following issue: “Do test results from one combined learning node represent all nodes within the combined node?”. Moreover, for evaluating the relation between learning effect and student ability, all participants are separated into three groups. The evaluation includes 115 participants. Participants are separated into a high-ability group, medium-ability group and low-ability group based on pre-test score. Both the high-ability and low-ability groups contain 31 participants, and the medium-ability group contains 53 participants. After all students finish their learning activities with LDDs, their LDDs and learning results are collected.

According to Table 1, CT represents three styles of Learning Diagnosis that are shown as Figs. 5a to 5c. The second row in the table represents combined learning nodes and the number of combined. For example, A(2) represents a new learning node combining two learning nodes. From Fig. 5a, learning node A is combined with “Resistance” and “Ohm’s law”. B(2) represents <Kirchoff’s law, Series connection>. From Fig. 5b, C(4) represents <Resistance, Ohm’s law, Kirchoff’s law, Series connection>. Moreover, D(2) represents <Diode, Zener diode>. According to Fig. 5c, E(8) represents <Electronic current, Resistance, Ohm’s law, Parallel connection, Kirchoff’s law, Series connection, Diode, Zener diode>. F(3) represents <Circuit, Thevenin Theorem, Wheaston Bridge>.

The bottom of the table contains three rows showing analytic results P(R) with high, medium and low ability students. The analytical results are calculated with formula $P(R) = P(B \mid A) = \frac{P(A \cap B)}{P(A)} X100\%$. P(A) represents the probability that students pass one combined learning node. Moreover, P(B|A) represents the probability of students passing one combined learning node, and then passing all learning nodes in the combined learning node. Higher P(R) represents that the combined and expanded algorithm in the adaptive LDD is effective.

Table 1 reveals that in the row with high-ability, all P(R) are larger than 90% except E(8). Because learning node E contains eight learning nodes, increasing P(R) means students have to simultaneously pass all eight learning nodes in learning node E. In the evaluation, although P(R) is not larger than 90%, 74% is also accurate.

In the column with CT=0.55, for three groups, all P(R) approach or exceed 80%. However, P(R) is between 83% and 42% for medium and low-ability groups with CT=0.5 and 0.45, meaning the two styles of the LDD are not suitable for these students.

The learning evaluation yields two important conclusions.

1. For high-ability groups with all styles of LDDs and with a number of combined learning nodes below 4, the degree of accuracy exceeds 90%.
2. For medium and low-ability groups with CT=0.55, the degree of accuracy approach or exceed 80%.

Table 1. Analytic results of P(R)

CT	0.55		0.5		0.45	
%	A(2)	B(2)	C(4)	D(2)	E(8)	F(3)
High	90	93	93	93	74	90
Medium	81	78	65	74	44	61
Low	80	83	62	59	42	77

5 Conclusions

In class teaching with a large number of students, the teachers are unable to spend much time on understanding individual student learning status. As a result, the fast-learning students thoroughly grasp the contents being taught in class, while the slow-learning students fall further and further behind, and eventually the education system gives up on them. In order to assist teachers to manage large classes, this research proposed a LDD system. The system provides personalized learning environment for individual students and adjusts learning phases to fit achievement of student automatically.

This research proposes a framework for remedial learning. In this research, we find out the LDD with different learning styles. In the learning evaluation, the adaptive LDD is suitable for high-ability students. Furthermore, the number of combined learning nodes cannot exceed four. Moreover, a suitable value for CT is between 1 and 0.5. Too small CT leads to too many learning nodes being combined, indicating decreased accuracy of the LDD test.

In the learning evaluation, the accuracy of CT=0.5 and 0.45 is not good enough for medium and low-ability students. There may be some combined factors that are not considered in the LDD. Possible such factors include the relationship between learning nodes and learning order. In the LDD, merging and expansion do not consider the relationship between learning nodes. Merging two learning nodes may cause combined learning nodes to mix with too much content or with very different content. The other factor is that the combined learning node does not consider learning order. When the evaluation course is being taught, two learning nodes may be separated into numerous weeks, but they are combined into a learning node that represents they learned simultaneously. In the future, research will be conducted on the two factors to increase the efficacy of the LDD.

References

1. Chang, F.C.I., Hung, L.P., Shih, T.K.: A New Courseware Diagram for Quantitative Measurement of Distance Learning Courses. Journal of Information Science and Engineering 19(6), 989–1014 (2003)
2. Goldsmith, T.E., Davenport, D.M.: Assessing structural similarity of graphs Pathfinder associative network: studies in knowledge organization, Norwood, pp. 75–87 (1990)
3. Goldsmith, T.E., Johnson, P.J., Acton, W.H.: Assessing Structural Knowledge. Journal of Educational Psychology 83(1), 88–96 (1991)
4. Hwang, G.J.: A conceptual map model for developing intelligent tutoring systems. Computers & Education 40(3), 217–235 (2002)
5. Jong, B.S., Chan, T.Y., Wu, Y.L.: Learning Log Explorer in E-learning Diagnosis. IEEE Transactions on Education 50(3), 216–228 (2007)
6. Jong, B.S., Lin, T.W., Wu, Y.L., Chan, T.Y.: An Efficient and Effective Progressive Cooperative Learning on the WEB. Journal of Information Science and Engineering 22(2), 425–446 (2006)
7. Jong, B.S., Wu, Y.L., Chan, T.Y.: Dynamic Grouping Strategies Based on a Conceptual Graph for Cooperative Learning. IEEE Transactions on Knowledge and Data Engineering 18(6), 738–747 (2006)
8. Novak, J.D., Goin, D.B.: Learning how to learn. Cambridge University Press, New York (1984)
9. Ruiz-Primo, M.A., Schultz, S.E., Li, M., Shavelson, R.J.: Comparison of the Reliability and Validity of Scores from Two concept-Mapping Techniques. Journal of Research In Science Teaching 38(2), 260–278 (2001)
10. Sowa, J.F.: Conceptual graphs for a data base interface. IBM Journal of Research and Development 20(4), 257–336 (1976)

New Integration Technology for Video Virtual Reality

Wei-Ming Yeh

Department of Radio & Television, National Taiwan University of Arts, Taipei, Taiwan 220
t0150@mail.ntua.edu.tw

Abstract. In Technical Image Press Association (TIPA) Awards 2009[1], Sony DSC-HX-1 received a honor of “Best Super zoom D-camera”. In fact, Sony HX-1 is recognized not only having a 20x optical super zoom lens, but offering many special video effects. It can be a new trend to judge current DSLR-like camera, such as Fujifilm FinePix HS10 received the same Award 2010[2] with same reasons again. Theoretically, it is a new integration technology for video virtually reality, which provide multiple platform users for video camera, video game, and mobile phone all together. Administers from variety of fields begin to think how to integrate some hot-selling video scene effects from all possible mobile video products, developing this “dazzling” virtually reality (VR) imagination beyond limitation, to attract more potential consumers, which can be vital for small businesses to survive in the future. In our experiment, we collect more than 300 cases from the telephone survey during March 2011 to Aug 2011. Total of 212 cases comply with the conditions. To probe mainly into the relationship between new generation video effects confidence level and 3 potential consumers: Amateur Photographer (AP), Senior Photographer (SP), and college student (CS). In our experiment, we collect more than 300 cases from the telephone survey during March 2011 to Aug 2011. Total of 212 cases comply with the conditions. To probe mainly into the relationship between new generation video effects confidence level and 3 potential consumers: Amateur Photographer (AP), Senior Photographer (SP), and college student (CS). That is the reason we are probe into this highly competitively market with brilliant creative design, and hope to offer an objective suggestion for both industry and education administers. .

Keywords: Augmented Reality, Cyber Codes, Fiduciary marker, HDR, Smart AR.

1 Introduction

Since early 2009, as soon as the digital video technology had great success in many fields, such as: video game, mobile phone, and digital video camera (DSC). Many small photo industries in Japan, realized that digital video technology could be the only way to compete with two giant photo tycoons (Canon and Nikon) in the market. In fact, Canon and Nikon have dominated the DSLR and SLR market for decades, and famous for high quality lens, CMOS and CCD sensor, processor, and accessories, not willing to follow this trend, to develop such fancy digital scene effect for their high-level DSLR camera surprisingly. Currently, Sony, M 4/3 family, and Samsung, realized that using digital video technology may achieve a niche between high-end DSLR, and consumer-level

J.-S. Pan, S.-M. Chen, N.T. Nguyen (Eds.): ACIIDS 2012, Part I, LNAI 7196, pp. 111–117, 2012.

DSC, offering shining stars, such as: Sony HX-1, HX-100V, Panasonic GF1, GF2, GF3, and many other products, recognized as prominent DSCs in TIPA 2009, 2010, 2011 Awards. It is a simple fact that DSLR-like camera and Mirrorless Interchangeable Lens Camera (MILC or EVIL) camera, with plenty of digital video technology (digital scene effects or customized scene effects), could attract many armature consumers and white-collar female consumers, offering giant marketing breakthrough, and survived in this margin-profit camera market. Actually, since 2007, the IT industries have co-operated with camera manufactures, developed many built in IC chips with brilliant customized scene effects for cameras (DSC mainly), such as: Face Detection Technology, Smile Shutter Mode, Full frame HD, CCD Anti Shake system, Live View, and many others. In fact, within few years, in early 2009, we found something new, such as: Back-illuminated CMOS image sensor[3], Sweep Panorama, Joining Multiple Exposed Patterns, Night Scene Portrait ,Motion Remover[7], that would challenge multimillion dollars business not in DC but highly profitable DSLR market, and may create a new possible attraction for photo-fans to replace their old camera.

In addition, since 2010, after the introduction of 3D movie (AVATAR), people all over the world seems to enjoy this 3D “mania” in their life, no matter 3D monitor for hardware, and 3D Panorama video effect for photo image., and Smart AR just another new technology wonder, which may create huge marketing benefit ahead. In fact, Smart AR is one of Sony new VR technology, which is related to Augmented reality (AR) , widely adopted by sports telecasting, video games, automobile, even jet-fighter weapon system(F-35), and ready for advanced video effect someday. It can offer a term for a live direct or an indirect view of a physical, real-world environment whose elements are augmented by computer-generated sensory input, such as sound or graphics. It is related to a more general concept called mediated reality, in which a view of reality is modified (possibly even diminished rather than augmented) by a computer. As a result, the technology functions by enhancing one’s current perception of reality. By contrast, virtual reality replaces the real world with a simulated one [5].

2 Technical Innovation

Generally, all hot-selling video scene effects are depending on the requirements from market. In the beginning of DSC, only expensive DSLR with high quality image and cheap compact could survived, since decades ago. As the progress of technology improvement, new categories of DSC were reveled, such as EVIL camera and DSLR-like, both equipped with many interesting video scene effects, to attract more consumers, those who enjoying private shooting, with neat size, and quality image, especially for white-collar female and low-key specialist. Therefore, a camera with plenty of built-in video scene effects, which is designed by many experienced photographers, can help people to take good picture effortless.

Currently, most armature consumers prefer to have a DSC with amazing, plenty video scene effects, no matter how often or how scarce using them, while shooting photos. People simply love it, and do create great profit. We notice some hot-selling items, which should be widely accepted and copied by many cameras, with reason price in 2009 till 2011, they were: blur-free twilight scenes shooting, Lightning-fast continuous shooting, Full HD Video, Sweep Panorama. As to another technical wonders, such as: 30× Optical

Zoom, Super High Speed ISO.., they might not be as popular (accepted)as people thought, or simply limited in few models. For example, in 2008 only a few DSLRs equipped with HD video. Currently, it is hardly to find a DSC without HD video ironically (before 2009). In addition, the Fujifilm HS10 (Fall 2009), offers two special scene effects: Motion Remover and Multi Motion Capture, which won a strong attention in this market, and HS 20 (Feb,, 2011) keep the good merit of Multi Motion Capture but delete Motion Remover(too complicate to handle). In addition, new Sony HX 100V (May, 2011) offers GPS, High Dynamic Range (HDR), and highly sophistic 3D Panorama, which may lead a new direction for future camera development. Coming with the new era of 3D technology, any specialist may take better 3D pictures easily, and manipulating 3D video image by special software, without bulky and expensive equipments. In addition, there are many brilliant 3D products available, such as: 3D TV, projector, Polarized 3D glasses, 3D game, and cell phone.

2.1 Phone Survey

In our experiment, we collect more than 300 cases from the telephone survey during March 2011 to Aug 2011. Total of 212 cases were effective, telephone surveys were written by two specialists, and those non-typical cases were determined after further discussion of 3 specialists.

2.2 Associated Analysis of Scene Effect Confidence Level

We took profile of contingency table proposed by Jobson (1992) to describe the preference level of scene effect associated with three user groups7. The study probes into preference confidence level and three specific groups, use chi-square test of independence to confirm the results.

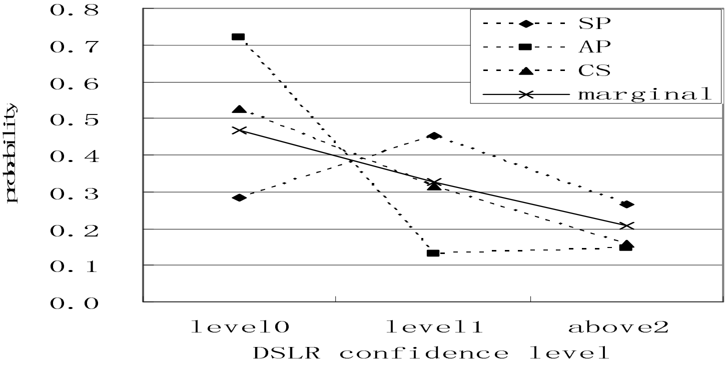
2.3 The Different Groups in Confidence Level

The relation of preference level and three groups we present as 3 × 4 contingency table (the preference level 0, 1, 2, 3 classified low, slight, moderate and high). Seldom have preference level "3", so we amalgamate "2" and "3" to "2" (above moderate) and then utilize this data to find out conditional and marginal probability, presented 3 × 3 probability table (Table.1).

Table 1. Group × Video effects Preference: Row proportions

Video effect	Preference level		
Group	0	1	2
SP	0.2830	0.4528	0.2642
AP	0.7206	0.1324	0.1471
CS	0.5263	0.3158	0.1579
marginal prob.	0.4670	0.3255	0.2075

Table 2. Group by Video effects preference profile



We utilize and draw profile of different groups with conditional and marginal probability of different groups preference level. It is for SP, AP and CS respectively in Fig. 3.Where solid line denotes the marginal probability of the levels, the dotted lines the conditional probability. Find levels "low"(level 0) and "slight"(level 1), there is a maximum disparity from solid line to others, nearly up to 0.2 in the AP, therefore can infer there is less preference level in the AP relatively, and greater probability of having slight confidence level in the SP, this shows it were to varying degrees influenced by the video effect preference, though with different categories of fan groups, hence the preference level associates with video effect. We use chi-squared test for independence to confirm and associated with the preference level, the -value would be 33.7058, and P =0.

3 Proportional Odds Model

We took a test for independence to confirm the different groups preference level and three groups in section 3: The different groups' preference level is related, but this test has not used the ordinal response level to confidence level. The regular collects the classification data with order in the camera research. For example it mentioned in the article the response variables of 'low', 'slight', 'moderate' and 'high' preference, we hope to probe into the factor influencing its preference level, so we analyze utilizing proportional odds model. if only one covariate, model present straight line relation to the logarithm of the accumulated odds of level p and covariate variable x, because this model supposes that there is the same slope β to all level p, this c-1 straight line is parallel each other. This model is based on McCullagh and Nelder theory (1989) and Diggle, Liang and Zeger (1994). One predictable variable was included in the study, representing known or potential risk factors. They are kinds of variable status is a categorical variable with three levels. It is represented by two indicator variables x (x_1 and x_2), as follows.

Table 3. 3 groups of indicator variables

kinds	x_1	x_2
SP	1	0
AP	0	1
CS	0	0

Note the use of the indicator variables as just explained for the categorical variable. The primary purpose of the study was to assess the strength of the association between each of the predictable variables. Let

$$L_p(x_1, x_2) = \theta_p + \beta_1 x_1 + \beta_2 x_2 , \quad p = 1,2 \tag{1}$$

Use likelihood ratio tests for $b_1 = b_2$ =0 hypothesis, the explanation of likelihood ratio tests is as follows: given L be the log likelihood of models, then G2= -2L. From hypothesis $b_1 = b_2$ =0, we have

$$L_p(x) = \theta_p , \quad p = 1,2,\cdots,c-1 \tag{2}$$

likelihood ratio tests use the difference of two deviances between model (1) and model (2) as reference value to test $H_0 : b_1 = b_2$ =0. If reject the hypothesis, furthermore, to test if b_1 =0 or b_2 =0. As fact, the preference level "under moderate" to the preference level "high", the odds ratio of SP compared with CS is also $\exp(b_1)$. Now according to formula, b_1 is the logarithm of the estimated odds when the preference level "under high" to the preference level "high", the SP compared with CS. b_1 >0 means the preference level presented by SP is less high than CS. Alternatively, b_1 <0 means the preference level presented by SP is higher than CS.

Therefore, $b_1 - b_2$ is the logarithm of the estimated odds when the preference level "under moderate" to the preference level "high", the SP compared with AP. $b_1 - b_2 > 0$ means the preference level presented by SP is less high than AP. Alternatively, $b_1 - b_2 < 0$ means the specify preference level presented by SP is higher than AP. On the rear part in the article, we will utilize odds ratio to probe into association between fan groups and video effects preference level.

3.1 Analysis of Different Groups Confidence Level

Combining confidence level "2" and " 3" to "2"(above moderate), we utilize odds ratio to calculate, deviance equal 416.96, if $b_1 = b_2 = 0$, deviance = 444.04, the difference between two deviances is 27.087, the related χ^2 -critical value will be test $H_0 : \beta_1 = \beta_2 = 0$, we find $\chi^2_{0.05} = 27.087$, and P=0.001. The null hypothesis is rejected and we conclude that β_1 and β_2 are not zero simultaneously. Furthermore, we analyze $\beta_1 = 0$ or $\beta_2 = 0$ or both not equal to zero. Table 3 is the results by maximum likelihood estimates. From Table 3 we find the hypothesis $\hat{\beta}_1 = 0$ is rejected, $\hat{\beta}_2 = 0$ is accepted, P-value is 0.017 and 0.072 respectively, $\hat{\beta}_2$ represents the logarithm of odds ratio for AP, thus the odds ratio would be estimated to be e^{β_2} . $\hat{\beta}_1 < 0$ represents the logarithm of odds ratio of preference≤ 0 rather than preference>0 of SP is 0.423 fold than that of CS. This indicates that the logarithm of odds ratio of preference>0 rather than preference≤ 0 of video effect preference Confidence level is about 2.36 fold for SP compared to CS customer. $\hat{\beta}_2 = 0$ means the confidence level presented by AP is the same as by CS, ascertaining the result is unanimous of this result.

Table 4. Analysis of the Video effect preference confidence level

Effect	B	Coefficient	Std Error	z-statistic	Sig
Intercept	$\hat{\theta}_1$	0.093	0.314	0.298	0.765
	$\hat{\theta}_2$	1.713	0.34	5.043	0.0000
Video effects	$\hat{\beta}_1$	-0.861	0.36	2.393	0.017
	$\hat{\beta}_2$	0.745	0.414	-1.802	0.072

3.2 Preference of Various Video Effects in Different Groups

Based on this data, AP (Amateur Photographer) may enjoy most of video effects, we assumed chip of video effects could save great amount of time and money, and pay more attention on “shooting”. As to Senior Photographer(SP), usually have professional software and better camera for years, they are famous to take quality picture by high-level DSLR with manual operation, doing post production with professional software, without any artificial intelligence effect. College student (CS) could have less budget and time to post small size photos for facebook or website frequently.

As to the 2011, the 3D Sweep Panorama (3DP), GPS, and High Dynamic Range (HDR), can be hot-selling scene effects now, many new DSCs already load them as part of standard equipments. We realize that Senior Photographer (SP) is willing to accept all possible new technologies and devices, and followed by Amateur Photographer (AP), as to College student (CS) show their conservative attitude all the times.

4 Conclusion

Since a decade ago, people have enjoyed the pleasure to take pictures or recording anything by various digital cameras (DSC/DV), in replace of old film-camera, and it can take good picture easily. The future of DSC will be full of surprised, handy in use, energy saving, affordable price, and more customized scene effects, in order to please photo-fans, video game users, which set a new record for successors already, and push camera and video manufactures to have more user-friendly innovation in entry-level model (Sony α series) especially. As to the latest wonders, Sony Smart AR combines the technology of 3D, Game, and VR, which may challenge the current hot-selling 3D panorama and VR effects in the future, and may create new 3D hardware market (DSC, Monitor/TV, Mobile Phone, NB..etc) soon. As to the future, this new technology can enhance one's current perception of reality image, and to simulate a new image, can be no limitation than ever.

References

1. TIPA Awards 2009: The best imaging products of 2009, TIPA (May 2010), `http://www.tipa.com/english/XIX_tipa_awards_2009.php`
2. TIPA Awards 2010: The best imaging products of 2010, TIPA (May 2011), `http://www.tipa.com/english/XX_tipa_awards_2010.php`
3. Rick User: New Sony Bionz processor, DC View (May 2, 2010), `http://forums.dpreview.com/forums/read.asp?forum=1035&message=35217115`
4. Alpha Sony: Sony Alpha 55 and 33 Translucent Mirror, Electronista (October 2010)
5. Graham-Rowe, D.: Sony Sets Its Sights on Augmented Reality, Technology Review (MIT) (May 31, 2011), `http://www.technologyreview.com/printer_friendly_article.aspx?id=37637`
6. Behrman, M.: This Augmented Reality Needs No Markers to Interfere With Your (Virtual) World, Gizmodo (May 19, 2011), `http://gizmodo.com/5803701/this-augmented-reality-needs-no-markers-to-interfere-with-your-virtual-world`
7. Moynihan, T.: Fujifilm's Motion Remover, PC World (March 3, 2010), `http://www.macworld.com/article/146896/2010/03/sonys_intelligent_sweep_panorama_mode.html`

Innovative Semantic Web Services for Next Generation Academic Electronic Library via Web 3.0 via Distributed Artificial Intelligence

Hai-Cheng Chu[1] and Szu-Wei Yang[2,*]

[1] Department of International Business
[2] Department of Education
National Taichung University of Education, 140 Min-Shen Road,
Taichung 40306 Taiwan ROC
{hcchu,swyang}@mail.ntcu.edu.tw

Abstract. The purpose of this article is to incubate the establishment of semantic web services for next generation academic electronic library via Web 3.0. The e-library users will be extremely beneficial from the forthcoming semantic social web mechanism. Web 3.0 will be the third generation of WWW and integrate semantics web, intelligent agent, and Distributed Artificial Intelligence into the ubiquitous networks. On top of current library 2.0 structures, we would be able to fulfill the Web 3.0 electronic library. We design the deployment of intelligent agents to form the semantic social web in order to interpret linguistic expressions of e-library users without ambiguity. This research is conducting the pioneering research to introduce the future and direction for the associate academic electronic library to follow the proposed guidelines to initiate the construction of future library system in terms of service-oriented architecture. This research article is the pioneering practice of future academic digital libraries under Web 3.0 structures.

Keywords: Web 3.0, Multi-Agent System (MAS), Digital Library, Semantic Social Web, Distributed Artificial Intelligence.

1 Introduction

The contemporary web technology has been focused on integration, virtualization, and socialization aspects (Cho, et al., 2008). Undoubtedly, Web 2.0 technology, which was emerged in 2000s, gradually dominates the current mainstreams of Internet Technology. Web 2.0 introduced the extraordinary phenomenon of user-generated contents or community-oriented gatherings based on the reality that it provides enormous users more than just retrieving information from existing web sites. It is constructed based upon architecture of participation that reduces the barrier of online collaboration and encourages the users to create and distribute the contents among communities. *Wikipedia*, *PLURK*,

[*] Corresponding author.

J.-S. Pan, S.-M. Chen, N.T. Nguyen (Eds.): ACIIDS 2012, Part I, LNAI 7196, pp. 118–124, 2012.

YouTube, *Flicker*, and *Twitter* are good examples of web 2.0 applications. Web 2.0 is the platform that people socialize or interact with each other and those people are brought together through a variety Community of Interest (COI) to form the cohesion. Any community member could share one’s sensation with the rest of the community participants either on-line or off-line. Some literatures suggest that Web 2.0 is also called Social Web, which strongly encourages community members to provide data or metadata in simpler ways regarding tagging, blogging, rating, or comments. Nevertheless, current Web 2.0 community web sites, substantial amount of data is scattered in the poorly structured databases and the contents are subjective prone. Hence, the repository of data dispersed around the networks is exponentially expanding. Unsurprisingly, finding the suitable data respecting a certain subject becomes much more challenging. Consequently, locating the appropriate and relevant information might be a tedious task for community members. The above arguments stimulate the emergence of Web 3.0, which will be the third wave of World Wide Web (WWW). The Intelligent Agent (IA) and Semantic Web play an essential role in the forthcoming Web 3.0 architectures. We have already witnessed library 2.0 serving the academic electronic library, which already focus on the SOA (Service-Oriented Architecture), from users’ perspectives [Yang, et al., 2009]. Till now, there is no academic electronic library that provides the Web 3.0 characteristics into the core functionalities of information systems. In this article, we propose the conceptual model of semantic web services for next generation academic electronic library via Web 3.0 architectures. Web 3.0 will have substantial effects regarding electronic library services in many aspects, which will dramatically change the ways for information search of digital libraries users.

2 Preliminaries

2.1 Intelligent Agent (IA)

The concept of IA was originated from the emergence of Distributed Artificial Intelligence (DAI), which was a major branch of AI, taken places around twenty years ago. Presently, it takes advantage of the ubiquitous networks and the power of distributed processing over the networks. An IA is regarded as a computer system, situated in some environment with the capability of flexible and autonomous actions in order to meet its design objective [Jennings, et al., 2008]. Hence, a Multi-Agent System (MAS) has proposed a new paradigm of mobile computing based on the cooperation and autonomy among them. Basically, one of the most challenging problems in a MAS is to ensure the autonomous characteristic of every single IA as a coherent behavior. Based on the related researches, social autonomy is believed to be one of the most important behaviors concerning the interactions between the agents in MAS [Carabelea, et al., 2004]. Social autonomy means the adoption of goals. From application point of view, the adoption of goals might exceed the processing capability of a certain IA. Under such circumstances, the IA can seek additional assistance from other IAs via the ubiquitous networks. In other words, an IA is able to delegate a goal to other IA, which might adopt it, through cooperation in order to solve the problems via the MAS structure. In some occasions, IA s might disagree concerning a specific goal. Hence, they might have to carry on negotiation in dynamic and pervasive computing environments. Additionally, there might be conflicts among the community of IAs. In this case, IAs are also capable of solving conflicts through negotiations. However, the details of the above frameworks or scenarios in MAS are beyond the scope of this research.

2.2 Web 3.0

Web 3.0 was a phrase coined by John Markoff of the New York Times in 2006. It refers to hypothetical Internet-based services that collectively integrate the related information that is randomly scattered in the networks and present to the end users without the intention to launch distinct application programs. Therefore, Web 3.0 is also called the Intelligent Web. Without loss of generality, Web 3.0 might be defined as the third-generation of the WWW enabled by the convergence of several key emerging technology trends encompassing ubiquitous connectivity, network computing, and the intelligent web [Nova, 2006]. Nowadays, Web 3.0 is already hotly debated among global Internet technology researchers. Although the definition of Web 3.0 is still indefinite, the reality is that many technologies and paradigms have been migrating into the prototype of this inevitable future. Some ICT (Internet Communication Technology) industry researchers and practitioners from different fields also indicate that Web 3.0 will ultimately be seen as applications that are pieced together and those applications would be very small and can be executed on any computing devices [Wikipedia, 2009].

Web 3.0 also refers to the stipulation of a more productive, personalized and perceptive surroundings for the end users by means of the integration of Semantic Web and Artificial Intelligence (AI) technologies, which acts as the core mechanism taking charge of interpreting linguistic expressions from the end users. Under Web 3.0 structures, all the information needs to be well-organized in order to be interpreted by machines without any ambiguity and as much as humans can. The Semantic Web will be bundled into the next generation of Web 2.0 and it will create web sites that can extract, share, re-use, and aggregate the information to one place presenting to users as collective knowledge. Consequently, the user interface is a crucial ingredient for the embracing of Web 3.0.

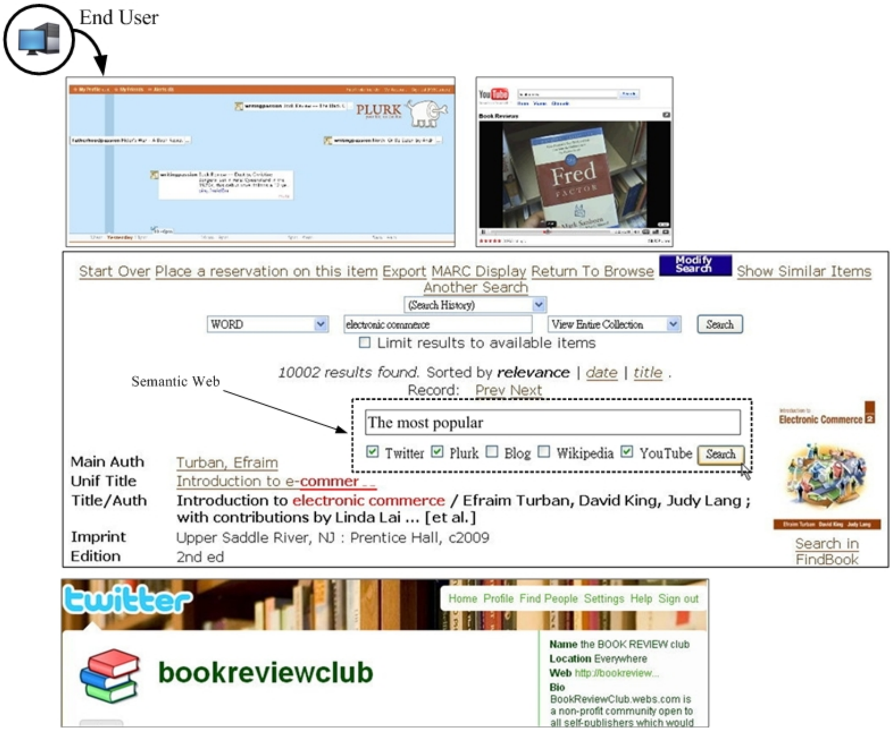
Notably, Web 3.0 will satisfy the users' needs in many aspects and it should be as natural as possible. The adoption of AI will provide collective intelligence and more collaborative searching power to the global academic digital library users. Under such circumstances, Web 3.0 will emphasize the advantages of semantic social web with respect to Web 2.0. Additionally, this will create the opportunities and possibilities for the use of semantic tagging and annotation for the social web. An IA that is embedded with the semantic web reasoning mechanism will search the web sites with the capability of processing complicated queries in order to retrieve the most optimal information for the users. Consequently, heterogeneous IAs are able to communicate, cooperate or even to proceed the conflict resolutions. Eventually, pervasive and ubiquitous computing will be the fundamental composites of Web 3.0, which will be omnipresent with multimodal intelligent user interfaces to access all kind of multimedia objects in an insensible manner. There is no doubt that for next generation of web services of digital libraries, the systems should be capable of providing end users more specific information based on semantic oriented applications.

At last, we would like to conclude that Web 3.0 is the combination of existing Web 2.0 and the Semantic Web. Ontology, IA, and semantic knowledge management will be integrated to the Semantic Web. Ambient Intelligence (AmI) will play an essential role concerning the embedded semantic reasoning mechanism in the third generation of WWW through heterogeneous networks. Furthermore, human machine interfaces will be much more unobtrusive to naive end users. Bringing high quality services to the users is a salient challenge of future academic electronic libraries. Although Web 3.0 is still in its infant age, several pioneering applications have touched down in the web communities. Those web sites emphasize on the application of AI, in other words, the more participants use it, the more convenient it becomes. They are powered by semantics interpretation and robotically learn the user preferences and would make connections and suitable recommendation tailored-made to end users. Those web sites will categorize distinct interests for users and eventually bring and bundle together pushing to users' desktop in one place and the users are able to share those with anyone.

The theme of this research paper is going to provide the system prototype of next generation academic library services under Web 3.0 structures, which none of the existing academic libraries have done. We also design and propose the conceptual model to illustrate the essence of the above statements from offering practical advices point of view. The paper incorporates the key elements of Web 3.0 into the core functionality for the web services of the future academic electronic libraries.

2.3 System Prototype of Web 3.0 Electronic Library

As Figure 1 illustrates, the dash rectangle area indicates the integration of Semantic Web into the current existing academic library search modules. As we know that many academic libraries subscribe electronic journals from publishers. Under Web 3.0 library structures, the user can provide the linguistic variables, which will be interpreted by the IA of a specific publisher. For example, the user found the desired book from current library information system. Furthermore, the user would like to screen related comments, reviews, or Q&A before check out the material. In our proposed scenario, the user might provide 'The most popular' in the dialog box as the linguistic expressions and check Twitter, PLURK, and YouTube as the reference resources as the figure illustrated. After the user submits the desired goal, the IA of the specific publisher will search the Q&A database within its own domain of the current publisher and provide the most optimal responses. Figure 2 depicts Web 3.0 library system prototype that we proposed. For example, academic library A subscribes Springer (electronic journals, eBooks, and Databases) and a certain user provides the goal through the dialog box designed in Figure 1. After the user submits the request, academic library A will notify Springer and the IA of the corporation will search the Q & A database within the publisher to provide the related responses. Concurrently, the IA will traverse to Web 2.0 community to compile and retrieve all the correlated information from Twitter, PLURK, Blog, Wikipedia, or YouTube. All the information will be delivered to the screen of the user who set the goal. The user would be able to visualize all the information without specifically triggering all those applications and this would be the beauty of Web 3.0 library. Figure 2 also demonstrates the scenarios that the IA of Springer might contact with the IA of Wiley, which is subscribed by academic library B instead of academic library A. MAS is formed to cooperate with each other even across heterogeneous information systems. In addition, the IA can cooperate across the publishers to aggregate the related knowledge and return to the user on the same screen. As figure 1 suggests, the IA will search the Web 2.0 community web sites, which were specified by the user and retrieve the information that the user desired and display those on the same screen without the user intentionally invoke the distinct applications. From the figures, it is obvious that the future library information will seamlessly couple with Web 2.0 with additional semantics reasoning functionalities embedded within an IA, which is capable of autonomously interpreting the linguistic expressions with accuracy in terms of information searching. Under such circumstances, the proposed Web 3.0 electronic library provides a new way for end users to collect online contents concerning comments, articles, Bolgs, photos, and videos in one place through parsing the linguistic expressions provided without the intentionally invoke specific application programs. Furthermore, the decision is upon the end users to share the information via RSS (Really Simple Syndication) mechanism. The future Web 3.0 electronic library will help end users collect information in a new, highly customized, and convenient way, which eventually makes the digital library web sites smarter.



electronic commerce

□ Limit results to available items

10002 results found. Sorted by relevance | date | title .

Record: Prev Next

The most popular

☑ Twitter ☑ Plurk □ Blog □ Wikipedia ☑ YouTube Search

Main Auth	Turban, Efraim
Unif Title	Introduction to e-commer
Title/Auth	Introduction to electronic commerce / Efraim Turban, David King, Judy Lang ; with contributions by Linda Lai ... [et al.]
Imprint	Upper Saddle River, NJ : Prentice Hall, c2009
Edition	2nd ed

Introduction to Electronic Commerce 2

Search in FindBook

Name the BOOK REVIEW club
Location Everywhere
Web http://bookreview...
Bio
BookReviewClub.webs.com is a non-profit community open to all self-publishers which would

Fig. 1. The search results that are automatically integrated presenting to end users

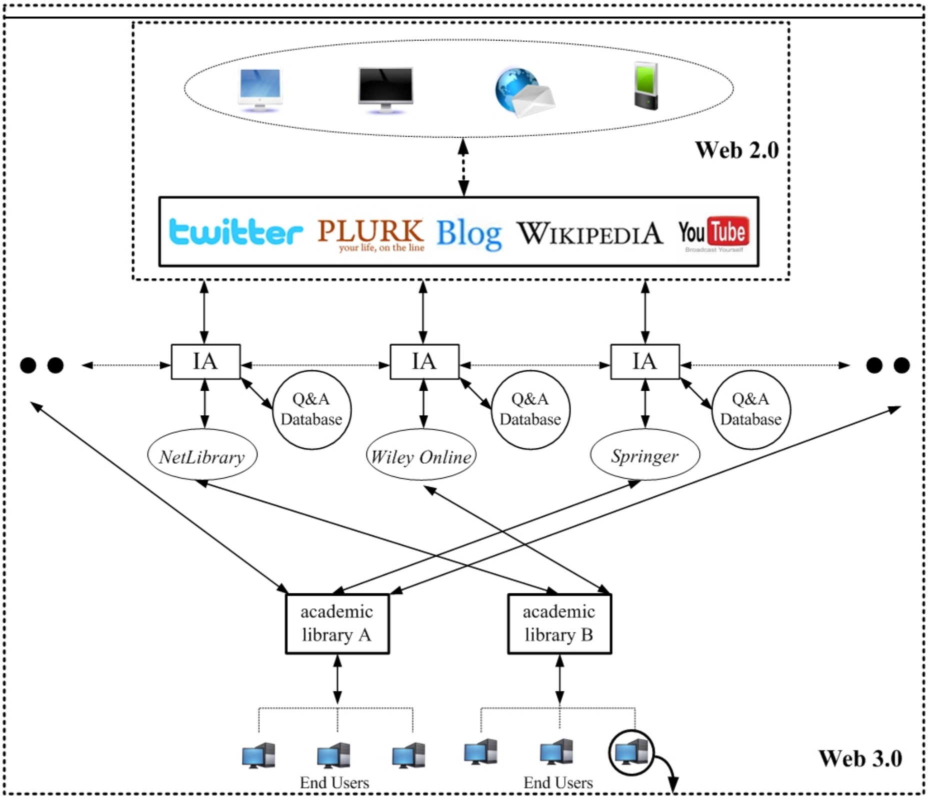


Fig. 2. Web 3.0 electronic libraries integrate semantic web, IA, and Web 2.0 community web sites

3 Conclusions

In this research, we propose the conceptual model of Web 3.0 digital library, which incorporates the Semantics Web mechanism into the core function of an IA. A web 3.0 library will be primarily based on pervasive and ubiquitous computing infrastructures. Currently, there is no academic electronic library fulfilling web 3.0 criteria. However, this is definitely the next generation of digital library. For any publisher, there will be an IA representing the publisher on the web and ready to cooperate or negotiate with other IAs in a MAS to fulfill the goal set by the users. The users would retrieve the related information without intentionally invoke the distinct applications and all those information would be collectively delivered to users in one place. Furthermore, automatic customization would be another feature of Web 3.0 due to the integration of AI. As technologies continue to advance, devices would become more context-aware and intelligent.

References

1. Carabelea, C., Boissier, O., Florea, A.: Autonomy in Multi-agent Systems: A Classification Attempt. In: Nickles, M., Rovatsos, M., Weiss, G. (eds.) AUTONOMY 2003. LNCS (LNAI), vol. 2969, pp. 103–113. Springer, Heidelberg (2004)
2. Cho, E.A., Moon, C.J., Park, D.H., Bait, D.K.: An Approach to Privacy Enhancement for Access Control Model in Web 3.0. In: Third International Conference on Convergence and Hybrid Information Technology, vol. 2, pp. 1046–1051 (2008)
3. Jennings, N.R., Sycara, K., Wooldridge, M.: A Roadmap of Agent Research and Development. International Journal of Autonomous Agents and Multi-Agent Systems 1(1), 7–38 (1998)
4. Silva, J.M., Rahman, A.S.M.M., Saddik, A.E.: Web 3.0: A vision for bridging the gap between real and virtual. In: Communicability MS 2008, Vancouver, BC, Canada, October 31, pp. 9–14 (2008)
5. Spivack, N.: The third-generation web is coming (2006), http://www.kurzweilai.net/articles/art0689.html?printable=1 (accessed on September 26, 2009)
6. Web 3.0 Definition, http://en.wikipedia.org/wiki/Semantic_Web (accessed on July 9, 2008)
7. Yang, X., Wei, Q., Peng, X.: System architecture of library 2.0. The Electronic Library 27(2), 283–291 (2009)
8. Santofimia, M.J., Fahlman, S.E., Moya, F., López, J.C.: A Common-Sense Planning Strategy for Ambient Intelligence. In: Setchi, R., Jordanov, I., Howlett, R.J., Jain, L.C. (eds.) KES 2010, Part II. LNCS (LNAI), vol. 6277, pp. 193–202. Springer, Heidelberg (2010)
9. Mamady, D., Tan, G., Toure, M.L., Alfawaer, Z.M.: An Artificial Immune System Based Multi-Agent Robotic Cooperation. In: Novel Algorithms and Techniques in Telecommunications, Automation and Industrial Electronics, pp. 60–67 (2008), doi:10.1007/978-1-4020-8737-0_12
10. Ford, A.J., Mulvehill, A.M.: Collaborating with Multiple Distributed Perspectives and Memories, Social Computing and Behavioral Modeling (2009), doi:10.1007/978-1-4419-0056-2_12
11. Mateo, R.M.A., Yoon, I., Lee, J.: Data-Mining Model Based on Multi-agent for the Intelligent Distributed Framework. In: Nguyen, N.T., Jo, G.-S., Howlett, R.J., Jain, L.C. (eds.) KES-AMSTA 2008. LNCS (LNAI), vol. 4953, pp. 753–762. Springer, Heidelberg (2008)
12. Chohra, A., Madani, K., Kanzari, D.: Fuzzy Cognitive and Social Negotiation Agent Strategy for Computational Collective Intelligence. In: Nguyen, N.T., Kowalczyk, R. (eds.) CCI I 2009. LNCS, vol. 6220, pp. 143–159. Springer, Heidelberg (2010)
13. Madani, K., Chohra, A., Bahrammirzaee, A., Kanzari, D.: SISINE: A Negotiation Training Dedicated Multi-Player Role-Playing Platform Using Artificial Intelligence Skills. In: Xhafa, F., Caballé, S., Abraham, A., Daradoumis, T., Juan Perez, A.A. (eds.) Computational Intelligence for Technology Enhanced Learning. SCI, vol. 273, pp. 169–194. Springer, Heidelberg (2010)
14. Daconta, M.C., Obrst, L.J., Smith, K.T.: The Semantic Web: A Guide to the Future of XML, Web Services, and Knowledge Management, 1st edn., pp. 3–26. Wiley (2003)
15. Scarlat, E., Maries, I.: Towards an Increase of Collective Intelligence within Organizations Using Trust and Reputation Models. In: Nguyen, N.T., Kowalczyk, R., Chen, S.-M. (eds.) ICCCI 2009. LNCS (LNAI), vol. 5796, pp. 140–151. Springer, Heidelberg (2009)
16. Breslin, J.C., Passant, A., Decker, S.: Towards the Social Semantic Web. In: The Social Semantic Web, pp. 271–283 (2009), doi:10.1007/978-3-642-01172-6_13
17. Boley, H., Osmun, T.M., Craig, B.L.: WellnessRules: A Web 3.0 Case Study in RuleML-Based Prolog-N3 Profile Interoperation. In: Governatori, G., Hall, J., Paschke, A. (eds.) RuleML 2009. LNCS, vol. 5858, pp. 43–52. Springer, Heidelberg (2009)

Using Fuzzy Reasoning Techniques and the Domain Ontology for Anti-Diabetic Drugs Recommendation

Shyi-Ming Chen[1,2], Yun-Hou Huang[1], Rung-Ching Chen[4],
Szu-Wei Yang[3], and Tian-Wei Sheu[2]

[1] Department of Computer Science and Information Engineering,
National Taiwan University of Science and Technology,
Taipei, Taiwan
[2] Graduate Institute of Educational Measurement and Statistics,
National Taichung University of Education, Taichung, Taiwan
[3] Department of Education, National Taichung University of Education, Taichung, Taiwan
[4] Department of Information Management, Chaoyang University of Technology,
Taichung, Taiwan

Abstract. In this paper, we use fuzzy reasoning techniques and the domain ontology for anti-diabetic drugs selection. We present an anti-diabetic drugs recommendation system based on fuzzy rules and the anti-diabetic drugs ontology to recommend the medicine and the medicine information. The experimental results show that the proposed anti-diabetic drugs recommendation system has a good performance for anti-diabetic drugs selection.

Keywords: Fuzzy Reasoning, Fuzzy Rules, Ontology, Anti-Diabetic Drugs.

1 Introduction

Clinical medicine expert systems have been presented for the past 40 years. The first generation of clinical medicine expert systems is the MYCIN system [11] which can diagnose infectious blood diseases. In recent years, some researchers [4], [5], [6], [8] combine the domain ontology with expert systems. The ontology techniques [1], [5]-[6], [8]-[9], are a combination of artificial intelligence and machine language to help to share and reuse the knowledge. It also contains natural language processing techniques and knowledge representation techniques. The ontology techniques can be used as channels of communication between human beings and systems. The ontology techniques can be further used for information retrieval and knowledge management. The more perfect the framework of domain ontology, the more complete the information which can be provided.

In this paper, we use fuzzy reasoning techniques and the domain ontology to build an anti-diabetic drugs recommendation system. The experimental results show that the proposed anti-diabetic drugs recommendation system has a good performance for anti-diabetic drugs selection.

J.-S. Pan, S.-M. Chen, N.T. Nguyen (Eds.): ACIIDS 2012, Part I, LNAI 7196, pp. 125–135, 2012.

2 Fuzzy Set Theory

In 1965, Zadeh proposed the theory of fuzzy sets [12]. Let X be a universe of discourse, where $X = \{x_1, x_2, ..., x_n\}$. A fuzzy set in the universe of discourse X can be represented as follows:

$$A = \sum_{i=1}^{n} \mu_A(x_i)/x_i \qquad (1)$$

$$=\mu_A(x_1)/x_1 + \mu_A(x_2)/x_2 + \ldots + \mu_A(x_n)/x_n\,,$$

where μ_A is the membership function of the fuzzy set A, $\mu_A(x_i)$ indicates the degree of membership of x_i in the fuzzy $\mu_A(x_i) \in [0, 1]$, the symbol ″+″ is the union operator, the symbol ″/″ is the separator, and $1 \leq i \leq n.$

Let A and B be two fuzzy sets in the universe of discourse U and let the membership functions of fuzzy sets A and B be μ_A and μ_B, respectively. Then, the union of the fuzzy sets A and B, denoted as $A \cup B,$ is defined as follows [12]:

$$\mu_{A\cup B}(u) = \max\{\mu_A(u), \mu_B(u)\}, \forall u \in U. \qquad (2)$$

The intersection of the fuzzy sets A and B, denoted as $A \cap B$, is defined as follows [2]:

$$\mu_{A\cap B}(u) = \min\{\mu_A(u), \mu_B(u)\}, \forall u \in U. \qquad (3)$$

3 Fuzzy Rules and Fuzzy Reasoning Techniques

Let us consider the following two fuzzy rules in the knowledge base of a fuzzy rule-based system:

IF X_1 is A_1 AND X_2 is B_1 THEN Z is C_1,
IF X_1 is A_2 AND X_2 is B_2 THEN Z is C_2,

where the observation is "X_1 is x and Y_1 is y" and x and y are crisp values. According to [3], Mamdni's Max-Min operations for fuzzy reasoning are shown in Fig. 1. Then, the fuzzy rule-based system performs the defuzzification operations to get a crisp value z of the fuzzy reasoning result based on the center of gravity (COG) defuzzification operations, shown as follows [3]:

$$z = \frac{\sum_{i=1}^{k} \mu_c(x_i)x_i}{\sum_{i=1}^{k} \mu_c(x_i)}, \qquad (4)$$

where μ_c is the membership function of the fuzzy set C, $\mu_c(x_i)$ denotes the degree of membership of x_i belonging to the fuzzy set C, $x_i \in$ w, and $1 \leq i \leq k.$

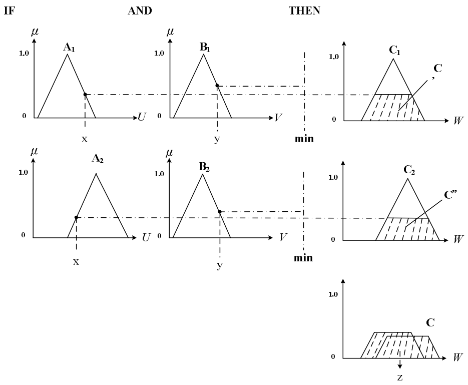


Fig. 1. Max-Min operations for fuzzy reasoning [3]

4 Ontology Knowledge

Ontology is a knowledge representation method in the semantic web [9] and it includes three parts, i.e., concepts, relationships and instances [5], [15]. In recent years, some researchers uses the Ontology Web Language (OWL) to describe the ontology. OWL is based on XML, where the RDF syntax is used in the OWL. The OWL can be divided into three levels of language [14], i.e., OWL Full, OWL DL and OWL List.

5 A Fuzzy Reasoning System for Recommending Anti-diabetic Drugs

In this section, we present a fuzzy reasoning system for recommending anti-diabetic drugs. Fig. 2 shows the structure of a fuzzy reasoning system for anti-diabetic drugs recommendation.

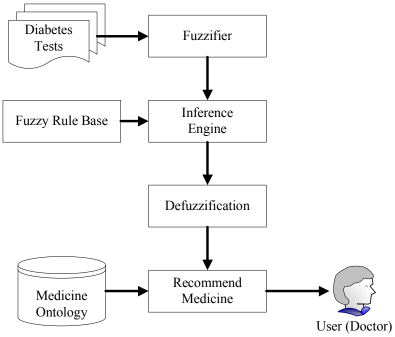


Fig. 2. A fuzzy reasoning system for recommending anti-diabetic drugs

In this paper, we adopt the clinical practice data of the American Association of Clinical Endocrinologists Medical Guidelines [10]. Table 1 shows a fuzzy rule matrix to infer the usability of the Metformin (MET) class anti-diabetic drugs, which contains 64 fuzzy rules with six attributes (HbA1c Test, Hypoglycemia Test, Renal Test, Heart Test, BMI Test and Liver Test) and three kinds of the usability, i.e., Recommend (R), Not Recommend (NR) and Danger (D); HbA1c is a test that measures the amount of glycated hemoglobin in the blood; the renal test is based on the Creatinine (Cr), which is a break-down product of creatine phosphate in the muscle and which is usually produced at a fairly constant rate by the body; the heart test is based on the functional classification of the New York Heart Association (NYHA); the NYHA functional classification provides a simple way of classifying the danger of heart failure; the weight test is based on the body mass index (BMI) or the Quetelet index, which is a heuristic proxy for the human body fat based on an individual's weight and height; the liver test is based on the liver's abnormal releases (GPT) [10]. Table 2 shows a fuzzy rule matrix to infer the usability of the Dipeptidyl peptidase-4 (DPP4) class anti-diabetic drugs. Table 3 shows a fuzzy rule matrix to infer the usability of the Thiazolidinedione (TZD) class anti-diabetic drugs. Table 4 shows a fuzzy rule matrix to infer the usability of the Glinide class anti-diabetic drugs. Table 5 shows a fuzzy rule matrix to infer the usability of the Sulfonylureas (SU) class anti-diabetic drugs. Table 6 shows a fuzzy rule matrix to infer the usability of the α-glucosidase (AGL) class anti-diabetic drugs to infer the degree.

Table 1. Fuzzy rule matrix to infer the usability of the Metformin class anti-diabetic drugs

		HbA1c	Normal				Abnormal			
		Hypoglycemia	No		Yes		No		Yes	
		Renal	Normal	Abnormal	Normal	Abnormal	Normal	Abnormal	Normal	Abnormal
Heart	BMI	Liver								
Normal	Low	Normal	R	D	R	D	R	D	R	D
		Abnormal	NR	D	NR	D	NR	D	NR	D
	High	Normal	R	D	R	D	R	D	R	D
		Abnormal	NR	D	NR	D	NR	D	NR	D
Abnormal	Low	Normal	R	D	R	D	R	D	R	D
		Abnormal	NR	D	NR	D	NR	D	NR	D
	High	Normal	R	D	R	D	R	D	R	D
		Abnormal	NR	D	NR	D	NR	D	NR	D

Table 2. Fuzzy rule matrix to infer the usability of the DPP4 class anti-diabetic drugs

		HbA1c	Normal				Abnormal			
		Hypoglycemia	No		Yes		No		Yes	
		Renal	Normal	Abnormal	Normal	Abnormal	Normal	Abnormal	Normal	Abnormal
Heart	BMI	Liver								
Normal	Low	Normal	R	R	R	R	R	R	R	R
		Abnormal	R	R	R	R	R	R	R	R
	High	Normal	R	R	R	R	R	R	R	R
		Abnormal	R	R	R	R	R	R	R	R
Abnormal	Low	Normal	R	R	R	R	R	R	R	R
		Abnormal	R	R	R	R	R	R	R	R
	High	Normal	R	R	R	R	R	R	R	R
		Abnormal	R	R	R	R	R	R	R	R

Table 3. Fuzzy rule matrix to infer the usability of the TZD class anti-diabetic drugs

		HbA1c	Normal				Abnormal			
		Hypoglycemia	No		Yes		No		Yes	
Heart	BMI	Liver / Renal	Normal	Abnormal	Normal	Abnormal	Normal	Abnormal	Normal	Abnormal
Normal	Low	Normal	R	R	R	R	R	R	R	R
		Abnormal	NR	NR	NR	NR	NR	NR	NR	NR
	High	Normal	NR	NR	NR	NR	NR	NR	NR	NR
		Abnormal	NR	NR	NR	NR	NR	NR	NR	NR
Abnormal	Low	Normal	D	D	D	D	D	D	D	D
		Abnormal	D	D	D	D	D	D	D	D
	High	Normal	D	D	D	D	D	D	D	D
		Abnormal	D	D	D	D	D	D	D	D

Table 4. Fuzzy rule matrix to infer the usability of the Glinide class anti-diabetic drugs

		HbA1c	Normal				Abnormal			
		Hypoglycemia	No		Yes		No		Yes	
Heart	BMI	Liver / Renal	Normal	Abnormal	Normal	Abnormal	Normal	Abnormal	Normal	Abnormal
Normal	Low	Normal	R	R	NR	NR	R	R	NR	NR
		Abnormal	NR	NR	NR	NR	NR	NR	NR	NR
	High	Normal	R	R	NR	NR	R	R	NR	NR
		Abnormal	NR	NR	NR	NR	NR	NR	NR	NR
Abnormal	Low	Normal	R	R	NR	NR	R	R	NR	NR
		Abnormal	NR	NR	NR	NR	NR	NR	NR	NR
	High	Normal	R	R	NR	NR	R	R	NR	NR
		Abnormal	NR	NR	NR	NR	NR	NR	NR	NR

Table 5. Fuzzy rule matrix to infer the usability of the SU class anti-diabetic drugs

		HbA1c	Normal				Abnormal			
		Hypoglycemia	No		Yes		No		Yes	
Heart	BMI	Liver / Renal	Normal	Abnormal	Normal	Abnormal	Normal	Abnormal	Normal	Abnormal
Normal	Low	Normal	R	NR	NR	NR	R	NR	NR	NR
		Abnormal	NR	NR	NR	NR	NR	NR	NR	NR
	High	Normal	R	NR	NR	NR	R	NR	NR	NR
		Abnormal	NR	NR	NR	NR	NR	NR	NR	NR
Abnormal	Low	Normal	R	NR	NR	NR	R	NR	NR	NR
		Abnormal	NR	NR	NR	NR	NR	NR	NR	NR
	High	Normal	R	NR	NR	NR	R	NR	NR	NR
		Abnormal	NR	NR	NR	NR	NR	NR	NR	NR

Table 6. Fuzzy rule matrix to infer the usability of the AGL class anti-diabetic drugs to infer the degree

		HbA1c	Normal				Abnormal			
		Hypoglycemia	No		Yes		No		Yes	
Heart	BMI	Liver / Renal	Normal	Abnormal	Normal	Abnormal	Normal	Abnormal	Normal	Abnormal
Normal	Low	Normal	R	NR	R	NR	R	NR	R	NR
		Abnormal	R	NR	R	NR	R	NR	R	NR
	High	Normal	R	NR	R	NR	R	NR	R	NR
		Abnormal	R	NR	R	NR	R	NR	R	NR
Abnormal	Low	Normal	R	NR	R	NR	R	NR	R	NR
		Abnormal	R	NR	R	NR	R	NR	R	NR
	High	Normal	R	NR	R	NR	R	NR	R	NR
		Abnormal	R	NR	R	NR	R	NR	R	NR

Fig. 3 shows the membership function curves for the HbA1c Tests, Hypoglycemia Test, Rental Test, Heart Test, BMI Test and Liver Test, respectively.

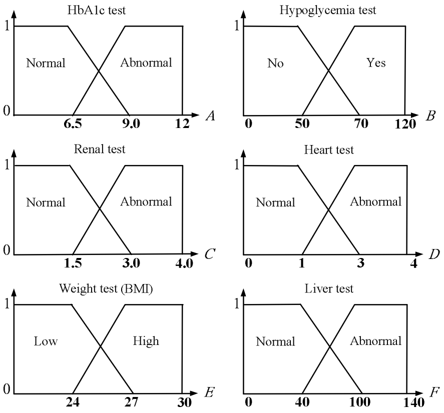


Fig. 3. Membership function curves for the tests

Fig. 4 shows the membership function curves for the fuzzy sets "Danger (D)", "Not Recommended (NR)", "Recommended (R)".

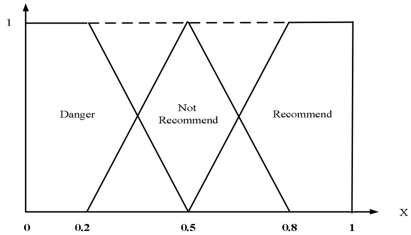


Fig. 4. Membership function curves for the fuzzy sets "Danger", "Not Recommended", and "Recommended"

A domain defines a set of representational knowledge that we call domain ontology. The domain represents the domain name of ontology and consists of various levels defined by domain experts. The domain ontology has three kinds of relationships, which are generalization, aggregation and association [4], [5], [6]. The generalization represents the "is-kind-of" relationship, which is a relationship between a domain and its corresponding category. The aggregation relationship is between each category and its corresponding events. The aggregation is a "is-part-of" relationship. The association represents a semantic relationship between concepts and concepts. Fig. 5 shows the structure of the anti-diabetes drugs knowledge. Diabetic drugs can be divided into six major medicine classes, including the α-glucosidase, the DDP4 inhibitor, the Glinide, the Metformin, the Sulfonylureas and the Thiazolidinedione. Thin subclasses includes the ingredient, the medicine name, the medicine doses and the medicine domain. In this paper, we combine fuzzy reasoning with anti-diabetic drugs ontology to recommend the best proposed medicine.

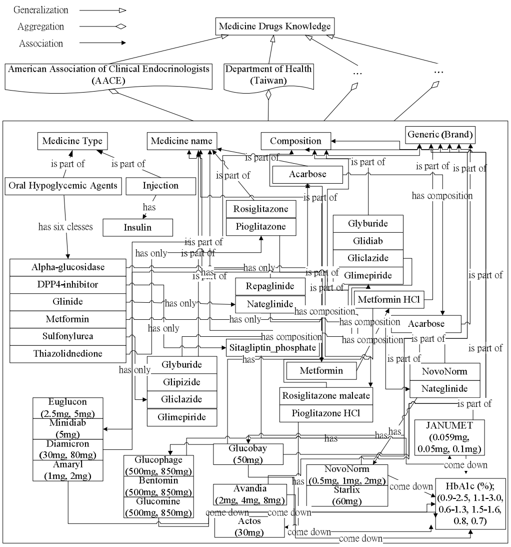


Fig. 5. Structure of the Anti-Diabetes Drugs Knowledge

6 Experimental Results

In this paper, we use the center of gravity method shown in formula (4) to deal with the defuzzification process. Table 7 shows 20 patients' data and their fuzzy reasoning results. Table 8 shows the 20 patients' data and their recommended levels. The recommended levels are 3, 2 or 1, which stand for different levels of recommendation. If the recommendation level is over 2, then the anti-diabetes drugs can be used. If the recommendation level is below 2, then the anti-diabetes drugs should be used carefully.

Table 7. 20 patients' data and fuzzy reasoning results

	HbA1c	Hypoglycemia	Renal	Heart	BMI	Liver	AGL	DPP4	Glinide	MET	SU	TZD
No.1	6.8	0	0.6	0	22	0	0.80837	0.80837	0.80837	0.80837	0.80837	0.80837
No.2	7.2	48	0.8	1	23	33	0.799297	0.799297	0.799297	0.799297	0.799297	0.799297
No.3	8.3	55	1.8	0	24.5	18	0.725865	0.799297	0.71293	0.658352	0.71293	0.735512
No.4	9.8	66	2.1	3	27	80	0.666831	0.791961	0.559477	0.483749	0.559477	0.208039
No.5	10	70	0.6	0	30	140	0.814551	0.814551	0.5	0.5	0.5	0.5
No.6	7.85	65	0.8	4	27	78	0.788143	0.788143	0.576484	0.605195	0.576484	0.211857
No.7	9	40	3.8	4	21	100	0.5	0.814551	0.5	0.235069	0.5	0.185449
No.8	11	65	3.9	2	25	78	0.5	0.794041	0.570937	0.242327	0.5	0.205959
No.9	6.5	60	1.5	3	26	80	0.785547	0.785547	0.601117	0.601117	0.601117	0.214453
No.10	11.5	68	2.7	2	24.8	130	0.554593	0.800086	0.5	0.272684	0.5	0.199914
No.11	7.9	0	2.1	4	23	100	0.660531	0.789427	0.5	0.389229	0.490959	0.210573
No.12	9.8	48	0.6	4	24.5	78	0.794041	0.794041	0.598304	0.598304	0.605367	0.205959
No.13	10	55	0.8	2	27	80	0.796089	0.796089	0.589085	0.589085	0.592071	0.203911
No.14	7.85	65	3.8	3	30	130	0.5	0.788143	0.5	0.235069	0.5	0.211857
No.15	9	60	3.9	2	27	98	0.5	0.785547	0.511751	0.242327	0.5	0.214453
No.16	6.75	40	3.9	3	30	18	0.5	0.80944	0.80944	0.242327	0.5	0.19056
No.17	6.9	65	1.5	0	27	80	0.796089	0.796089	0.569414	0.589085	0.569414	0.5
No.18	7.78	60	2.7	4	21	140	0.5654	0.785547	0.5	0.303666	0.5	0.214453
No.19	6.8	68	2.1	4	25	78	0.666831	0.791961	0.530927	0.491988	0.530927	0.208039
No.20	6.65	0	0.6	2	26	100	0.796089	0.796089	0.5	0.5	0.391938	0.203911

Note: The range of "Recommended" is between 0.65 and 1, the "Not Recommended" is between 0.35 and 0.65, and the range of "Danger" is between 0 and 0.35.

Table 8. 20 patients' data and recommendation levels

	HbA1c	Hyppoglycemia	Rennal	Heart	BMI	Liver	AGL	DPP4	Glinide	MET	SU	TZD
No.1	6.8	0	0.6	0	22	0	3	3	3	3	3	3
No.2	7.2	48	0.8	1	23	33	3	3	3	3	3	3
No.3	8.3	55	1.8	0	24.5	18	3	3	3	3	3	3
No.4	9.8	66	2.1	3	27	80	3	3	2	2	2	1
No.5	10	70	0.6	0	30	140	3	3	2	2	2	2
No.6	7.85	65	0.8	4	27	78	3	3	2	2	2	1
No.7	9	40	3.8	4	21	100	2	3	2	1	2	1
No.8	11	65	3.9	2	25	78	2	3	2	1	2	1
No.9	6.5	60	1.5	3	26	80	3	3	2	2	2	1
No.10	11.5	68	2.7	2	24.8	130	2	3	2	1	2	1
No.11	7.9	0	2.1	4	23	100	3	3	2	2	2	1
No.12	9.8	48	0.6	4	24.5	78	3	3	2	2	2	1
No.13	10	55	0.8	2	27	80	3	3	2	2	2	1
No.14	7.85	65	3.8	3	30	130	2	3	2	1	2	1
No.15	9	60	3.9	2	27	98	2	3	2	1	2	1
No.16	6.75	40	3.9	3	30	18	2	3	3	1	2	1
No.17	6.9	65	1.5	0	27	80	3	3	2	2	2	2
No.18	7.78	60	2.7	4	21	140	2	3	2	1	2	1
No.19	6.8	68	2.1	4	25	78	3	3	2	2	2	1
No.20	6.65	0	0.6	2	26	100	3	3	2	2	2	1

Note: "3" denotes "Recommended", "2" denotes "not recommended", and "1" denotes "Danger".

Table 9 [10] shows the medicine names and their composition for reducing the HbA1c. We construct the ontology to remind the oral hypoglycemic agents' knowledge. Protégé was used to construct the medicine ontology in the preliminary experiment due to the fact that the Protégé supports the RDF (Resource Description Framework). The OWL DL (Description Logic) format [14] was adopted in this paper due to the fact that the OWL DL is based on the XML and RDF syntax. OWL DL supports those users who want the maximum expressiveness while retaining the computational completeness (i.e., all conclusions are guaranteed to be computable) and decidability (i.e., all computations will be finished in a finite time) [14].

Table 9. Oral hypoglycemic drugs knowledge [10]

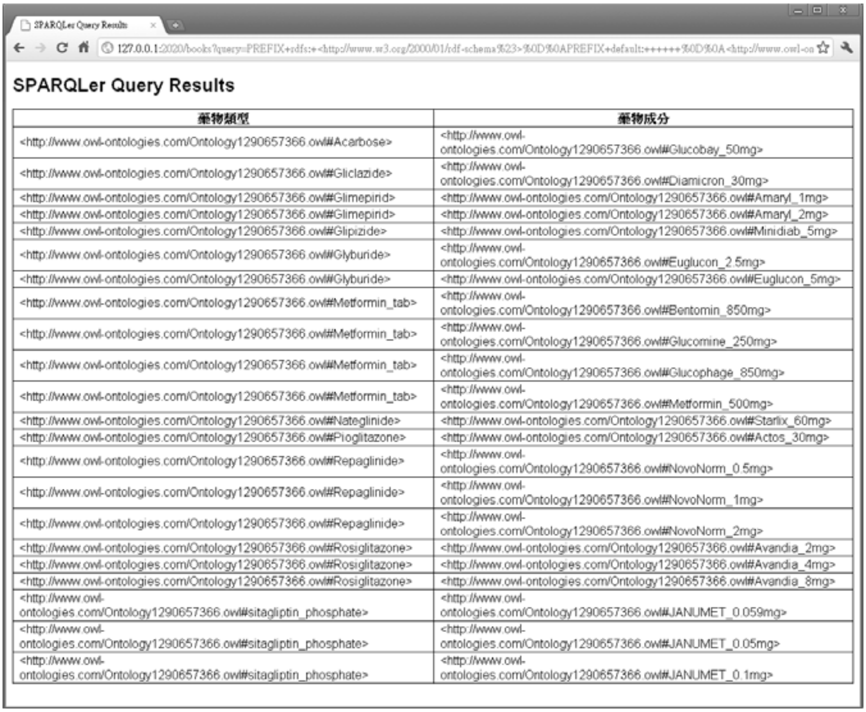
Medicine Class	Medicine Name / Generic Name, Brand Name	Composition	HbA1c (%)
Sulfonylureas (SU)	Glyburide / Euglucon	Glyburide (2.5mg, 5mg)	0.9 - 2.5
	Glipizide / Minidiab	Glidiab (5mg)	
	Gliclazide / Diamicron	Gliclazide (30mg, 80mg)	
	Glimepiride / Amaryl	Glimepiride (1mg, 2mg)	
Metformin (MET)	Metformin / Glucophage, Bentomin, Glucomine	Metformin HCl (500 mg, 850 mg)	1.1 - 3.0
α-glucosidase (AGL)	Acarbose / Glucobay	Acarbose (50mg)	0.6 - 1.3
Thiazolidinedione (TZD)	Rosiglitazone / Avandia	Rosiglitazone maleate (2mg, 4mg, 8mg)	1.5 - 1.6
	Pioglitazone / Actos	Pioglitazone HCl (30mg)	
Glinide	Repaglinide / NovoNorm	NovoNorm (0.5mg, 1mg, 2mg)	0.8
	Nateglinide / Starlix	Nateglinide (60mg)	
DDP4 inhibitor	Sitagliptin_phosphate	JANUMET (0.059mg, 0.05mg, 0.1mg)	0.7

The anti-diabetic drugs ontology created is by Protégé. “Classes”, “Properties” and “individuals” are used to describe the data. The class of medicine, attribute of medicine class and contents of medicine properties are set up to construct the relevant knowledge of anti-diabetic drugs.

In this paper, we use Joseki to implement the web server, which uses SPARQL to query the knowledge base, such as the user input diabetic drugs (Metformin), and then the system returns the ingredient and the medicine doses. Joseki [13] is an HTTP engine which supports the SPARQL Protocol and the SPARQL RDF Query language. Joseki Features has the RDF Data from files and databases and the HTTP (i.e., GET or POST) implementation of the SPARQL protocol. SPARQL was developed from W3C RDF Data Access Working Group. The SPARQL Protocol and RDF Query Language (SPARQL) is a query language and the protocol for RDF [16].

The format of the SPARQL Query language is as follows:

```
PREFIX rdfs: <http://www.w3.org/2000/01/rdf-schema#>
PREFIX default:
<http://www.owl-ontologies.com/Ontology1290657366.owl#>
SELECT ?藥物類型 ?藥物成分
WHERE
{ ?藥物類型 default:names ?藥物成分 }
ORDER BY ASC (?藥物類型) ASC(?藥物成分)
```


SPARQLer Query Results
127.0.0.1:2020/books?query=PREFIX+rdfs:+<http://www.w3.org/2000/01/rdf-schema%23>%0D%0APREFIX+default:++++++%0D%0A<http://www.owl-on

SPARQLer Query Results

藥物類型	藥物成分
<http://www.owl-ontologies.com/Ontology1290657366.owl#Acarbose>	<http://www.owl-ontologies.com/Ontology1290657366.owl#Glucobay_50mg>
<http://www.owl-ontologies.com/Ontology1290657366.owl#Gliclazide>	<http://www.owl-ontologies.com/Ontology1290657366.owl#Diamicron_30mg>
<http://www.owl-ontologies.com/Ontology1290657366.owl#Glimepirid>	<http://www.owl-ontologies.com/Ontology1290657366.owl#Amaryl_1mg>
<http://www.owl-ontologies.com/Ontology1290657366.owl#Glimepirid>	<http://www.owl-ontologies.com/Ontology1290657366.owl#Amaryl_2mg>
<http://www.owl-ontologies.com/Ontology1290657366.owl#Glipizide>	<http://www.owl-ontologies.com/Ontology1290657366.owl#Minidiab_5mg>
<http://www.owl-ontologies.com/Ontology1290657366.owl#Glyburide>	<http://www.owl-ontologies.com/Ontology1290657366.owl#Euglucon_2.5mg>
<http://www.owl-ontologies.com/Ontology1290657366.owl#Glyburide>	<http://www.owl-ontologies.com/Ontology1290657366.owl#Euglucon_5mg>
<http://www.owl-ontologies.com/Ontology1290657366.owl#Metformin_tab>	<http://www.owl-ontologies.com/Ontology1290657366.owl#Bentomin_850mg>
<http://www.owl-ontologies.com/Ontology1290657366.owl#Metformin_tab>	<http://www.owl-ontologies.com/Ontology1290657366.owl#Glucomine_250mg>
<http://www.owl-ontologies.com/Ontology1290657366.owl#Metformin_tab>	<http://www.owl-ontologies.com/Ontology1290657366.owl#Glucophage_850mg>
<http://www.owl-ontologies.com/Ontology1290657366.owl#Metformin_tab>	<http://www.owl-ontologies.com/Ontology1290657366.owl#Metformin_500mg>
<http://www.owl-ontologies.com/Ontology1290657366.owl#Nateglinide>	<http://www.owl-ontologies.com/Ontology1290657366.owl#Starlix_60mg>
<http://www.owl-ontologies.com/Ontology1290657366.owl#Pioglitazone>	<http://www.owl-ontologies.com/Ontology1290657366.owl#Actos_30mg>
<http://www.owl-ontologies.com/Ontology1290657366.owl#Repaglinide>	<http://www.owl-ontologies.com/Ontology1290657366.owl#NovoNorm_0.5mg>
<http://www.owl-ontologies.com/Ontology1290657366.owl#Repaglinide>	<http://www.owl-ontologies.com/Ontology1290657366.owl#NovoNorm_1mg>
<http://www.owl-ontologies.com/Ontology1290657366.owl#Repaglinide>	<http://www.owl-ontologies.com/Ontology1290657366.owl#NovoNorm_2mg>
<http://www.owl-ontologies.com/Ontology1290657366.owl#Rosiglitazone>	<http://www.owl-ontologies.com/Ontology1290657366.owl#Avandia_2mg>
<http://www.owl-ontologies.com/Ontology1290657366.owl#Rosiglitazone>	<http://www.owl-ontologies.com/Ontology1290657366.owl#Avandia_4mg>
<http://www.owl-ontologies.com/Ontology1290657366.owl#Rosiglitazone>	<http://www.owl-ontologies.com/Ontology1290657366.owl#Avandia_8mg>
<http://www.owl-ontologies.com/Ontology1290657366.owl#sitagliptin_phosphate>	<http://www.owl-ontologies.com/Ontology1290657366.owl#JANUMET_0.059mg>
<http://www.owl-ontologies.com/Ontology1290657366.owl#sitagliptin_phosphate>	<http://www.owl-ontologies.com/Ontology1290657366.owl#JANUMET_0.05mg>
<http://www.owl-ontologies.com/Ontology1290657366.owl#sitagliptin_phosphate>	<http://www.owl-ontologies.com/Ontology1290657366.owl#JANUMET_0.1mg>

Fig. 6. SPARQL Query Result

Fig. 6 shows the SPARQL query results regarding the ingredient and medicine doses of diabetic drugs. For example, the user query the medicine composition, the Acarbose, the Gliclazide, …, and the Sitagliptin phosphate. The system returns medicine dose, i.e., Glucobay: 50mg, Diamicorn: 30mg, …, and JANUMET: 0.1mg.

7 Conclusions

In this paper, we have used fuzzy reasoning techniques and the domain ontology for anti-diabetic drugs selection. We have presented an anti-diabetic drugs recommendation system based on fuzzy rules and the anti-diabetic drugs ontology to recommend the medicine and the medicine information. The experimental results show that the proposed anti-diabetic drugs recommendation system has a good performance for anti-diabetic drugs selection.

Acknowledgment. The authors would like to thank Dr. Cho-Tsan Bau, Taichung Hospital, Taichung, Taiwan, for his help during this work. This work was supported in part by the National Science Council, Republic of China, under Grant NSC100-2221-E-011-118-MY2.

References

1. Bobillo, F., Delgado, M., Gómez-Romero, J., López, E.: A Semantic Fuzzy Expert System for a Fuzzy Balanced Scorecard. Expert Systems with Applications 36(1), 423–433 (2009)
2. Chen, S.M., Lee, S.H., Lee, C.H.: A New Method for Generating Fuzzy Rules from Numerical Data for Handling Classification Problems. Applied Artificial Intelligence 15(7), 645–664 (2001)
3. Lee, C.C.: Fuzzy Logic in Control Systems: Fuzzy Logic Controller, Part II. IEEE Transactions on Systems, Man, and Cybernetics 20(2), 419–435 (1990)
4. Lee, C.S., Wang, M.H.: A Fuzzy Expert System for Diabetes Decision Support Application. IEEE Transactions on Systems, Man, and Cybernetics, Part B: Cybernetics 41(1), 139–153 (2011)
5. Lee, C.S., Wang, M.H., Hagras, H.: A Type-2 Fuzzy Ontology and Its Application to Personal Diabetic-Diet Recommendation. IEEE Transactions on Fuzzy Systems 18(2), 374–395 (2010)
6. Lee, C.S., Jian, Z.W., Huang, L.K.: A Fuzzy Ontology and Its Application to News Summarization. IEEE Transactions on Systems, Man, and Cybernetics-Part B: Cybernetics 35(5), 859–880 (2005)
7. Misra, S., Roy, S., Obaidat, M.S., Mohanta, D.: A Fuzzy Logic-Based Energy Efficient Packet Loss Preventive Routing Protocol. In: Proceedings of the 12th International Conference on Symposium on Performance Evaluation of Computer & Telecommunication Systems, SPECTS 2009, pp. 185–192 (2009)
8. Mao, Y., Wu, Z., Tian, W., Jiang, X., Cheung, W.K.: Dynamic Sub-Ontology Evolution for Traditional Chinese Medicine Web Ontology. Journal of Biomedical Informatics 41(5), 790–805 (2008)
9. Quan, T.T., Hui, S.C., Fong, A.C.M.: Automatic Fuzzy Ontology Generation for Semantic Help-Desk Support. IEEE Transactions on Industrial Informatics 2(3), 1551–3203 (2006)

10. Rodbard, H.W., Blonde, L., Braithwaite, S.S., Brett, E.M., Cobin, R.H., Handelsman, Y., Hellman, R., Jellinger, P.S., Jovanovic, L.G., Levy, P., Mechanick, J.I., Zangeneh, F.: American Association of Clinical Endocrinologists Medical Guidelines for Clinical Practice for the Management of Diabetes Mellitus. American Association of Clinical Endocrinologists 13, 1–68 (2007)
11. Shortliffe, E.H.: MYCIN: A Rule-Based Computer Program for Advising Physicians Regarding Antimicrobial Therapy Selection. Technical Report, Department of Computer Sciences, Stanford University, California (1974)
12. Zadeh, L.A.: Fuzzy Sets. Information and Control 8, 338–353 (1965)
13. Joseki - A SPARQL Server for Jena, `http://joseki.sourceforge.net/`
14. OWL Web Ontology Language Overview, `http://www.w3.org/TR/owl-features`
15. Ontology, `http://en.wikipedia.org/wiki/Ontology`
16. SPARQL, `http://www.w3.org/TR/rdf-sparql-protocol/`

Content-Aware Image Resizing Based on Aesthetic

Jia-Shing Sheu, Yi-Ching Kao, and Hao Chu

Department of Computer Science, National Taipei University of Education
jiashing@tea.ntue.edu.tw

Abstract. This paper presents an image resizing system, the purpose is to build an image resizing system based on aesthetic composition including rule of thirds and subject position. With global operations, using the traditional scaling and non-traditional content-aware image resizing do the global operation, supplemented by photo rating for adaptive adjustment to reduce user operation with clear quantitative criteria as a basis for adjustment. In non-traditional content-aware image resizing, two algorithms are used, Seam Carving for Content-Aware Image Resizing and Adaptive Content-Aware Image Resizing, to adjust path detection with Otsu's Method according to the above-mentioned algorithm diagram.

Keywords: Photo composition, Esthetics rule, Content-Aware Image Resizing, Otsu’s Method, Photo Rating.

1 Introduction

Nowadays, the popularity of digital cameras allows people take picture everywhere, but not every photo can be satisfactory, so more and more image modification software has be develop, but general public are not professional photographers nor photo retouching experts, the public want to modify the composition of their favorite photos to the photographer used even meet the aesthetic point of view, so effective and simple to use software to adjust the image has become a trend by the attention. The focus of this paper is to reduce user operations, and bring out the traditional image adjustment and the two non-traditional method of image adjustment, in addition the subjective factors in the observed images and the background are the main things twisted and deformed, and factors in the objective for using a different platform assessment scores before and after image adjustment, used to justify what was better, and integration of different scoring system in order to achieve adaptive image adjustment purposes.

2 Literature Review

Each person has a different aesthetics definition, the so-called aesthetic composition can also be said to be very subjective, but the rule of thumb is that it can be integrated

J.-S. Pan, S.-M. Chen, N.T. Nguyen (Eds.): ACIIDS 2012, Part I, LNAI 7196, pp. 136–145, 2012.

out and for a publically recognized standard, Hong-Cheng Kao proposed [1], several criteria for the aesthetic image under the composition is well defined and measurable criteria, and in Che-Hua Yeh and others have proposed [2], they extend the original composition standard and developed a set of evaluation and ranking system that can be adopted by users for their shoots. The photos selected in this study were analyzed by the system, too.

However, in terms of photo resizing, manual resizing is deeply depended due to different individual subjective aesthetic. In 2007, what Shai Avidan has presented [3] can perceive the photo contents to further avoid the main object and overall algorithm when resizing photo composition that can reduce resizing. The next year, 2008 Michael Rubinstein presented [4] made a very good extension application, but in the process of photo resizing have some defects. Such as the protection of the contents is still not enough or the adjustment direction is not satisfactory. For this problem, Yu-Shuen Wang using visual saliency improvement image energy map [5]. For [3] the choice of path and improve efficiency, Wei-Ming Dong [6] also made some improvements.

In this paper, an image resizing system is proposed, using content-aware image resizing [3] [6]. In addition, we give a threshold for the energy map, because these algorithm path choices bases on energy map [7]. The goal is to smoothen the smooth part and reduce the main object distortions. In aesthetic composition adaptation, we tried traditional scaling to collocate with non-traditional content perception that allows the resized photos more accepted to the general public.

3 Image Adjustment System Architecture

3.1 System Overview

First, use Photo Ranking System [2] to give a score for the original image by the three elements on composition. Then through our system choose algorithm to resizing. After the adjustment is completed, then by [1][2] to give the different aspects scores of subjective and objective. In order to adjust the image does have consistent aesthetic basis, as shown in Fig. 1 at Sec. 3.2. In this system, Esthetic Analysis background is mentioned in [1][2], Content Aware Image Resizing[3][6] is coupled with threshold selection method [7] to adjust the weights of the energy which mapped to compare and analyze.

3.2 System Flow Chart

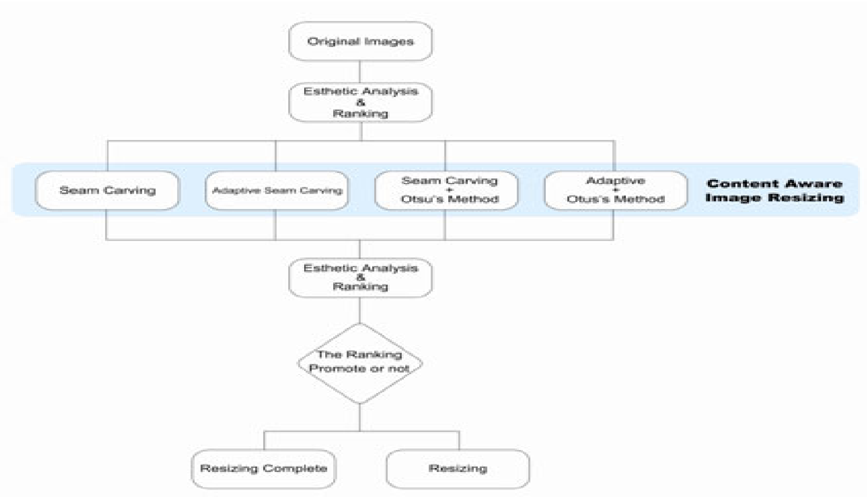

Original Images
Esthetic Analysis & Ranking
Seam Carving
Adaptive Seam Carving
Seam Carving + Otsu's Method
Adaptive + Otus's Method
Content Aware Image Resizing
Esthetic Analysis & Ranking
The Ranking Promote or not
Resizing Complete
Resizing

Fig. 1. Image adjustment system

3.3 The Basic Theory of Experimental Design

3.3.1 Image Resizing Based on the Rule of Thirds

"Golden ratio" by the ancient Greeks invented geometry formulas, following this rule of composition is deemed to be "harmonious". When enjoying an image or an artwork, this rule would provide a reasonably separated Geometric segment. For many painters and artists, the "golden ratio" is that they must thoroughly understand the creation of a guiding principle; it is also widely used in photographic composition. The rule of thirds mentioned in [1], is simplified from the "golden ratio", (1:1.6), and its basic purpose is to avoid symmetrical composition. With symmetrical composition, the subject matter is usually placed in the center of the frame, often making the photo look dull and monotonous. However, the horizontal line used to divide the image into one-third and two-thirds of the image so that it no longer confined to the monotony of the symmetrical composition, to increase the picture's proportionality, as Fig. 2.

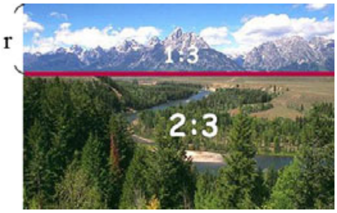


Fig. 2. Rule of thirds example

The chosen images detecting the horizon by the algorithm mention in [1]. which r is distance from the horizon to the upper.

DH is the Horizontal line for the Rule of Thirds in the range of composition.

${}^{\sigma}H_r$ is the standard deviation of the Rule of Thirds.

${}^{s}H_r$ is the original score of the horizontal line, and final score for the Rule of Thirds is ${}^{s}H_b$.

If $DH_{1.6} > r > DH_{2.0}$, then the score ${}^{s}H_b$ is 1.0. However r not in this region, then the ${}^{s}H_b$ is the maximum absolute value from the three equations in of ${}^{S}H_r$, and ${}^{S}H_b$ in this region. The result of ${}^{S}H_r$ is closer to 1.0, representing the horizon closer to the Rule of Thirds. When the score is 1.0, represents horizon match the rule of thirds.

$$D = \min[r, (Image\ height - r)] \tag{1}$$

$$DH_{1.0} = [Image\ height]/(1+1.0) \tag{2}$$

$$DH_{1.6} = [Image\ height]/(1+1.6) \tag{3}$$

$$DH_{2.0} = [Image\ height]/(1+2.0) \tag{4}$$

$${}^{\sigma}H_r = \frac{(DH_{1.6} - DH_{2.0})}{2} \tag{5}$$

$${}^{s}H_{r=1.0} = -\exp\left(\frac{-(D-DH_{1.0})^2}{2{}^{\sigma}H_r{}^2}\right) \tag{6}$$

$${}^{s}H_{r=1.6} = -\exp\left(\frac{-(D-DH_{1.6})^2}{2{}^{\sigma}H_r{}^2}\right) \tag{7}$$

$${}^{s}H_{r=2.0} = -\exp\left(\frac{-(D-DH_{2.0})^2}{2{}^{\sigma}H_r{}^2}\right) \tag{8}$$

Use these equation can calculate the score for the image. Magnitude of each adjustment of system is one-tenth of the image high. Here it is defined DR. $DR = [Image\ height]/10)$, To do for the [3] [6] algorithm adjust each time the image pixels. Each time after adjustment by the algorithm is re-calculated scores, and when the score is 1.0, the adjustments are finished.

3.3.2 Image Resizing Based on the Subject Location

Basis on horizontal and vertical the Rule of Thirds, can find the intersection - Power Weight Point. This means that subject position is better on these four golden points.

Chosen image that main subject is deviation from Power Weight Point, use [3] [6] algorithm to resizing, each time adjustment by DR. Use f_{ROT} in [2] to score the subject and four Power Weight Point. A_i is the subject size. S_i is the saliency value of subject. D_i is the distance of subject and Power Weight Point(standard deviation $\sigma = 0.17$). If the subject is more close to the Power Weight Point scores will be higher.

$$f_{ROT} = \frac{1}{\sum_i A_i S_i} \sum_i A_i S_i e^{-\frac{D_i^2}{2\sigma}} \tag{9}$$

After the image resizing base on the rule of thirds and the subject relocation, need to be observed the background and subject have been destroyed or distorted.

3.3.3 Binary Threshold Otsu's Method

Otsu's method [7] can determination a binary threshold value t. Gray scale images can be represented as a 2D gray level functions, including N pixel, gray scale values between K to L, the pixels have the same gray value i is f_i. The probability of the gray value image i emersion in image can be expressed as P_i. In bi-level binary Threshold case, pixels will be divided into two categories, C_1 is gray scale value of pixels $[K, \ldots, t]$, and C_2 is gray scale value of pixels $[t+1, \ldots, L]$. $\omega_1(t)$ is the probability of gray scale value of pixels K to t total probability. $\omega_2(t)$ is the probability of gray scale value of pixels $t+1$ to L total probability. C_1 and C_2 with the average of the whole image to gray scale the average grayscale value of whole image to represent by μ_1, μ_2 and μ_T. Otsu's method [7] definition the between-class variance can use σ_B^2 equation to solution. Otsu's method find the best threshold value should in the grayscale value region between K to L. T is Otsu's method threshold value.

$$P_i = fi / N \tag{10}$$

$$\begin{aligned} &C_1 : P_k / \omega_k(t), \ldots P_t / \omega_1(t) \text{ and} \\ &C_2 : P_{t+2} / \omega_2(t), P_{t+2} / \omega_2(t), \ldots P_L / \omega_2(t) \\ &\text{where } \omega_2(t) = \sum_{i=k}^{t} P_i,\ \omega_2(t) = \sum_{i=i+1}^{L} P_i \end{aligned} \tag{11}$$

$$\mu_1 = \sum_{i=k}^{t} P_i / \omega_1(t)\ ,\ \mu_2 = \sum_{i=i+1}^{L} P_i / \omega_2(t) \tag{12}$$

$$\omega_1 \mu_1 + \omega_2 \mu_2 = \mu_T \tag{13}$$

$$\sigma_B^2 = \omega_1(\mu_1 - \mu_T)^2 + \omega_2(\mu_2 - \mu_T)^2 \tag{14}$$

$$T = \underset{K<t<L}{Arg\ \ Max}\{\sigma_B^2(t)\} \tag{15}$$

3.4 Image Resizing Method - Non-traditional Image Resizing Method Fusion the Binary Threshold Otsu's Method

Image through aesthetic analysis, resizing for different parts of the content is used the non-traditional Content-Aware Image Resizing [3] [6] and Scaling in traditional way, for the [3] [6] In particular, we addition Otsu's Method on the path selection method is also Energy map to adjustment.

The main aim is use the Sobel Edge Detection on the original content-aware resizing of to detect the energy map. $e(I)$

$$e(I) = \left|\frac{\partial}{\partial x} I\right| + \left|\frac{\partial}{\partial y} I\right| \tag{16}$$

However, this approach is not use Binary Threshold on the Energy map $e(I)$, the optimal threshold value T for the Energy map is calculated.

Were substituted into the [3][4], Zero below the T part and keep the value higher than T :

$$\begin{aligned} if\ \ & e(i,j) \le T, e(i,j) = 0 \\ & e(i,j) > T, e(i,j) = e(i,j) \end{aligned} \tag{17}$$

Make Content-Aware Image Resizing in the path choice can easily avoid the subject and select smooth part is the path. While the energy overall sum of Seam selected (the energy due to some zero on the path) will different; the Optimal Seam will be different. So the smooth part allowed to more easily identify the more is our main purpose.

After we resizing the image, score the image by [1] [2] respectively, and observe the background and the main object distortion and destruction, which is used to judge the resized image which one is better.

4 Experiments

4.1 Image Resizing Results

We use the Image base on Rule of Thirds and the subject location with different coping strategies to score and analysis. First choices the image have horizon line on the middle. Then use non-traditional content-aware resizing algorithms to adjust. According the experiments we give energy map a threshold or not, to adjust energy map of the different results.

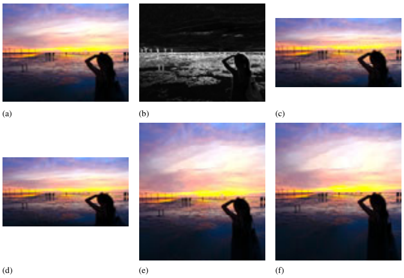


Fig. 3. Adjusted the horizon to the rule of thirds and the threshold value *T* is not set
(a) Original Image: The horizon on the image middle.
(b) Original Image`s energy distribution.
(c) Reduce pixels by Seam Carving resizing to adjust horizon.
(d) Reduce pixels by Adaptive to adjust horizon.
(e) Increase pixels by Seam Carving resizing to adjust horizon.
(f) Increase pixels by Adaptive to adjust horizon.

From the Fig. 3, we can observe that, using [3] [6] algorithm to make adjustments and the energy map with threshold value not set Horizontal line in the adjustment process cannot successfully adjust to the rule of thirds. Because the right-down silhouette energy too closes the background energy, adjustment process chose to adjust right-down silhouette. It cannot achieve the purpose we want - to adjust the horizontal line.

The original energy map threshold *T* is given. Then adjust and observe. In Fig. 4 can see the clouds at top on picture become smoother after using the threshold. Adjust by original [3][6] algorithm will adjust the silhouette like Fig. 3. After give threshold vale, algorithm will adjust clouds. Then can adjust to the rule of thirds on the horizontal line.

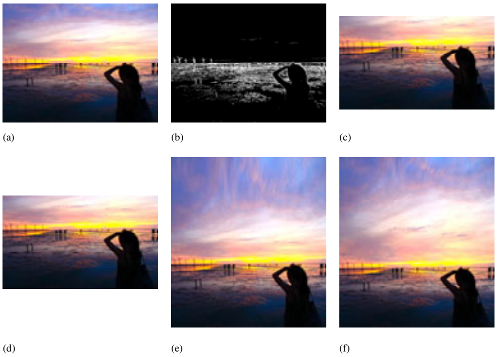


Fig. 4. Adjusted the horizon to the rule of thirds and the threshold value *T* is set
(a) Original Image: The horizon on the image middle.
(b) Original Image`s energy distribution by give the threshold value *T*.
(c) Reduce pixels by Seam Carving resizing to adjust horizon.
(d) Reduce pixels by Adaptive to adjust horizon.
(e) Increase pixels by Seam Carving resizing to adjust horizon.
(f) Increase pixels by Adaptive to adjust horizon.

At subject location adjustment, we choose a picture that subject on image center. Using the above method to give the threshold value *T* for the [3][6] algorithm used for comparison.

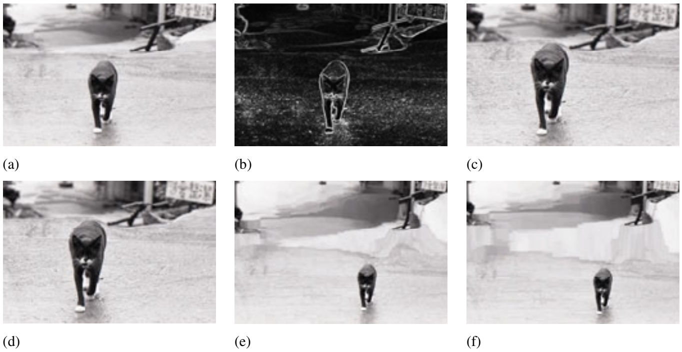


Fig. 5. The subject adjusts to Power Weight Point and the threshold value is not set
(a) Original Image: The subject in the center.
(b) Original Image`s energy distribution
(c) Reduce pixels by Seam Carving resizing to adjust subject location.
(d) Reduce pixels by Adaptive to adjust subject location.
(e) Increase pixels by Seam Carving resizing to adjust subject location.
(f) Increase pixels by Adaptive to adjust subject location.

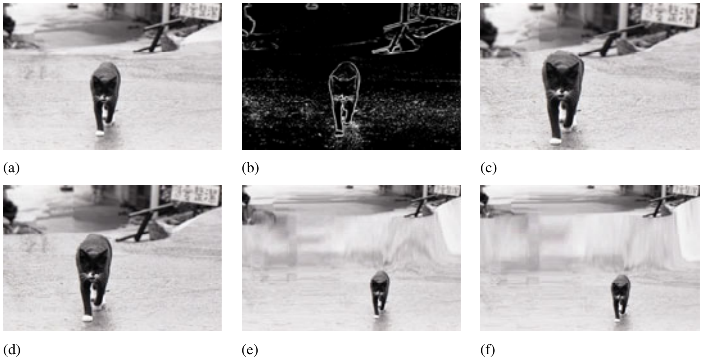


Fig. 6. The subject adjusts to Power Weight Point and the threshold value is set
(a)Original Image: The subject in the center.
(b) Original Image`s energy distribution by give the threshold value *T*.
(c) Reduce pixels by Seam Carving resizing to adjust subject location.
(d) Reduce pixels by Adaptive to adjust subject location.
(e) Increase pixels by Seam Carving resizing to adjust subject location.
(f) Increase pixels by Adaptive to adjust subject location.

From Fig. 5 and Fig. 6 can see because the energy of subject higher than background, so whether the given threshold for the adjustment process of the subject re-location to little effect. On the process of adjustment pixels Increase or decrease the pixel process, decrease pixel is better than increase pixel. Decrease pixel processes have less frequent operations. The Increase pixel process to cause amplification background is too vague or subject led to the amplification of the background over the main image as is the proportion of the overall decline.

4.2 Image Adjustment Score Rating

Adjust the horizontal line in Fig. 3, after the horizontal line near the middle. The results of ${}^{s}H_{b}$ are close to the worst case -1, and adjusted Fig. 4, the score of the horizontal line close to 1.0.

In Fig. 5 and Fig. 6 is not very different, the form below do the presentation and comparison.

Table 1. Fig. 5 and Fig. 6s' f_{ROT}

Fig. 5	f_{ROT}	Fig. 6	f_{ROT}
(a)	0.120294	(a)	0.120294
(b)	N/A	(b)	N/A
(c)	0.25625	(c)	0.348914
(d)	0.37424	(d)	0.348457
(e)	0.241147	(e)	0.107835

f_{ROT} is ranking system scores. Both of Fig. 5 and Fig. 6, (b) the energy map for the image not graded. By the table above we can clearly see that image scores adjust subject location by reduce the image pixels (c) (d) improved significantly than the unadjusted image (a) scores. Increase pixels to adjust subject location (e) (f), the f_{ROT} increase magnitude are smaller. And, Fig. 6 (given threshold T) compare with Fig. 5 (not given threshold T), Fig. 6 (c) score is higher than Fig. 5 (c). That is, [3] algorithm has been improvement by [7] given threshold. But in (e) and (f) contrary to the scores is declined. In the same we found that the scores of subject re-location by reduce pixels (c) and (d) higher than Increase pixels (e) (f). Adjust subject location by Increase pixels will cause background over-amplification and fuzzy. Thus a conclusion will be given, adjust the position of the subject to the four Power Weight Point, use [3] and [6] method with reduce pixels is better way, and if choose [3] to adjust, can give threshold T [7] to increase the performance.

5 Conclusion

According the experiments, in horizon adjustment, the subject energy and background energy too close (Fig. 5, 6 (a) Original Image), use content-aware resizing to do horizon adjust often damage to the subject image distortion and the horizon line. But after given threshold to reduce the non-subject energy, content-aware resizing can be more easily to select background to be path.

Seam Carving for Content-Aware Image Resizing with threshold can improve the results than original Seam Carving for Content-Aware Image Resizing. Adaptive Content-Aware Image Resizing is a less obvious effect. Finally, use reduce pixels than increase pixel is better, reason is that the increase pixel process to cause amplification background is too vague or subject led to the amplification of the background over the main image as is the proportion of the overall decline.

The above study found that after a given threshold, Content-Aware Image Resizing algorithms can protection the subject. Adjusted to meet the aesthetic composition in the different images have different adjustment methods. The system for different types of images with traditional and non-traditional methods allowed more convenient to adjust and perfect, and we hope that the future can make the system more efficient development of more complete, so that each image can be easily adjusted to meet the aesthetic composition, each individuals can have a photographer shoot out like a beautiful image.

References

1. Kao, H.-C.: Esthetics-based Quantitative Analysis of Photo Composition. Master's Thesis, Department Computer Science & information Engineering, National Taiwan University, R.O.C (2008)
2. Yeh, C.-H., Ho, Y.-C., Barsky, B.A., Ouhyoung, M.: Personalized Photograph Ranking and Selection System. In: Proceedings of the International Conference on Multimedia (2010)
3. Avidan, S., Shamir, A.: Seam Carving for Content-Aware Image Resizing. ACM Transactions on Graphics (New York, NY, USA) 26(3) (July 2007)
4. Rubinstein, M., Shamir, A., Avidan, S.: Improved Seam Carving for Video Retargeting. ACM Transactions on Graphics (TOG) - Proceedings of ACM SIGGRAPH (New York, NY, USA) 27(3) (August 2008)
5. Wang, Y.-S., Tai, C.-L., Sorkine, O., Lee, T.-Y.: Optimized Scale-and-Stretch for Image Resizing. ACM Transactions on Graphics (TOG) - Proceedings of ACM SIGGRAPH Asia 27(5) (December 2008)
6. Dong, W., Paul, J.-C.: Transactions on Adaptive Content-Aware Image Resizing, vol. 28(2). The Eurographics Association and Blackwell Publishing Ltd. (2009)
7. Otsu, N.: A Threshold Selection Method. IEEE Transactions on System, Man, and Cybernetics, 62–66 (1979)

A Functional Knowledge Model and Application

Nhon Do and Thu-Le Pham

Department of Computer Science, Information Technology University,
Ho Chi Minh City, Vietnam
{nhondv,thuplta}@uit.edu.vn

Abstract. In artificial intelligence, knowledge models play an important role in designing the knowledge base systems and the expert systems. The quality of the intelligent systems depends heavily on the knowledge base built on the models. This paper presents a model of knowledge called functional knowledge model (FKM). This model is used to represent knowledge domains about functions and computing relations between them in different real applications. Besides, this paper also proposes a technique for solving an important problem in functional knowledge domains - simplification of functional expressions. Functional knowledge model is applied to construct an automatic system for simplifying trigonometric expressions in high school. This system can reason automatically from knowledge and provides a step by step solution that similar to the way of human thinking.

Keywords: Artificial intelligence, knowledge representation, educational software, automatic problem solving.

1 Introduction

Knowledge representation is one of important branches of artificial intelligence. To build the expert systems or knowledge-based problem solving systems, we need to design a knowledge base and an inference engine. The quality of the intelligent systems depends greatly on knowledge. Therefore, researching and developing methods for representing knowledge and reasoning mechanisms have great significance in theories and applications in artificial intelligence science.

Nowadays, there are many different knowledge models have been proposed and are widely applied in many fields. The knowledge models are mentioned in [1], [2], [5], [10] such as semantic network, rule – based system, conceptual graph, computing network, COKB, etc, for human many different methods in designing the knowledge base.

The components of knowledge, which are represented by those models, are rather varied such as: concepts, operators, laws, etc. However, in human knowledge, there are many other components have not been studied fully for representing. One of those components is functions and their computing relations. Therefore, in this paper, we present a knowledge model for representing a domain related functions and their relations. This model is a method for designing a functional knowledge base for solving automatically problems of inference and calculation on functions. Simplification is a basic and important problem in this knowledge domain. This paper also proposes a

J.-S. Pan, S.-M. Chen, N.T. Nguyen (Eds.): ACIIDS 2012, Part I, LNAI 7196, pp. 146–155, 2012.

technique for automatic solving this problem by means of defining the length of expressions.

One of the most practical applications of this model is knowledge domain of trigonometric expressions. Automated and readable simplification of trigonometric expressions is an important issue but has not been completely solved by current computer algebra systems. In [11], the author presented a technique for solving this problem by means of combination rules. However, this technique can simplify trigonometric polynomials only. Therefore, based on proposed model and algorithm, the paper also presents the design of a program that simplifies trigonometric expressions automatically. The program is implemented by Maple and C #. This program can simplify not only trigonometric polynomials but also other forms of trigonometric expressions. It is able to give result and solution in step by step for explaining the process of reasoning. It is very useful for teachers and students in teaching and studying.

Functional knowledge model is a contribution in the research and development of methods for representing knowledge on the computer.

2 A Model of Functional Knowledge

The knowledge models in [1], [2], [5], [10] has been widely used in many practical applications. However, these methods are not suitable and enough for representing the knowledge domains of functions on computers. The following functional knowledge model is a useful method for designing the knowledge base of functional concepts, operators and their relations.

2.1 Functional Knowledge Model (FKM)

FKM is a knowledge model consists of 4 components:

(Cf, Ops, Rules, Vals)

in which,

- Cf: is a set of functional concepts.
- Ops: is a set of operators.
- Rules: is a set of rules. A rule is a computing relation on functions that represented by a functional equation.
- Vals: is a set of specific values of functions.

Notation: <Cf> is a set of functional expressions that built by functions and operators in Ops.

A functional concept is a class of functions. The structure of a function consists of:

- Function kind,
- A set of variables. A variable has a kind of variable.
- A set of parameters. A parameter has a kind of parameter.
- Kind of result of function.
- A set of constrains of parameters.
- Definition of function

Example 1: Logarithm to the base *a* of *x* has structure:

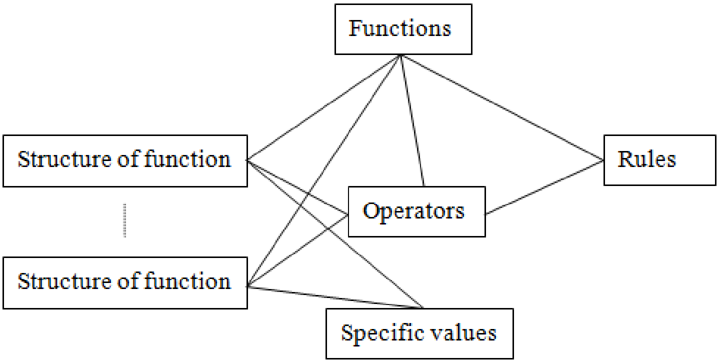
Function kind: LOGARIT
Variable: *x*, variable kind: REAL
Parameter: *a*, parameter kind: REAL
Result kind: REAL
Constrains: $a > 0, a \neq 1$
Function definition: $\log[\mathrm{a}](\mathrm{x})$.

2.2 Storage Organization

The knowledge base of functions is represented by FKM can be organized by a system of structured text files. They include the files below:

- The file "Functions.txt": stores name of functional concepts.
- The file "Operators.txt": stores information of operators.
- The file "Rules.txt": stores computing rules.
- The file "Values.txt": stores specific values of functions.
- The files "<name of concept>.txt": stores the specifications of structures of functions.

The relationship of information structure in the knowledge base can be illustrated by the following diagram:



The relationship of components in FKM

3 Problem and Algorithms

In this section, an important problem in functional knowledge domain is mentioned. It is simplification of functional expressions. The problem is stated as:

Given functional knowledge model FKM (Cf, Ops, Rules, Vals), and an expression exprf ∈ *<Cf>. Simplify exprf.*

In other words, to simplify expression exprf is to transform exprf into a new expression exprh with exprh "is simpler than" exprf. In the process of simplification, we know that this process will stop when we cannot make expression simpler any more. But how do we teach computers to know that? Therefore, to simplify automatically expressions on computers, we need to define the concept of "simpler than" explicitly. This concept is a key to build algorithms of simplification.

Here are rules for computing length of expression exprf in <Cf> (Notation: length(exprf))

(1). *exprf* is a constant function then *length(exprf) = 0.*

(2). *exprf* is a function in functional class in Cf and *exprf* is not a constant function then *length(exprf) = 1.*

(3). *exprf* = *exprg*^{n} *, which exprg* ∈ *<Cf> then*

- *length(exprf) = |n|*length(exprg) if n ∈ Z*
- *length(exprf) = length(exprg) if n ∈ Q/Z*

(4). *exprf = exprg*exprh* then *length(exprf) = length(exprg)*length(exprh)*, with *exprg, exprh ∈ <Cf>,* and operator ** ∈ Ops.*

Example 2: $f = \dfrac{\sin(x)}{\sin(x) - \sqrt{\cot(x)^2 - \cos(x)^2}} - \dfrac{\cos(x)^2}{2\sin(x)^2 - 1}$ ⇒ length(f) = 10.

3.1 Definitions

Here are some definitions that are necessary for the design of automatic simplification algorithm. Given FKM (Cf, Ops, Rules, Vals):

Definition 1. Given two expressions exprf, exprg ∈ <Cf>. exprf is ”simpler than” exprg if length(exprf) < length(exprg).

Definition 2. Given an expression *exprf* ∈ <Cf>. A rule r ∈ Rules is “applied” to *exprf* if left – hand side of r is a sub expression of *exprf.*

Definition 3.

– A rule r in Rules is called “simplifying rule” if $length(lhs(r)) > length(rhs(r))$. lhs(r), rhs(r) are left – hand side and right – hand side of r.

A set of simplifying rules is Rrg.

$$Rrg = \{r \in Rules, length(lhs(t)) > length(rhs(r))\}$$

– A rule in Rules is called “expanding rule” if $length(lhs(r)) \le length(rhs(r))$.

A set of expanding rules is Rkt.

$$Rkt = \{r \in Rules, length(lhs(t)) \le length(rhs(r))\}$$

Base on definition 3, the set Rules in FMK consists of Rrg and Rkt.

$$Rules = Rrg \cup Rkt$$

3.2 Algorithms

The general simplification algorithm below is based on the definition mentioned in section 3.1. The objective of simplifying expression is to transform an original expression into a new expression whose length is smaller. The process of simplification is not always possible to reduce the length. It is necessary to make a combination of simplifying and expanding to achieve the desired results. General simplification algorithm below is the combination of two processes: the "simple simplification process" and the "simple expansion process” until achieve the expression has the shortest length. The process of "simple simplification" is the process of finding rules in Rrg with some basic transformation methods such as common factor, etc, to create a new expression has shorter length. The process of "simple expansion" is the process of finding rules in Rkt with some basic transformation methods such as: transform product into sum, common denominator, etc.

The General Simplification Algorithm

```
Input: an expression exprf
Output: an expression exprg is a result of simplification of exprf
    Step 1: Simplify simply exprf and set up initial values for variables.
    exprg is a result of simple simplification process of exprf. exprg_old
    stores expression whose length is smallest in each step.
        exprg:= simple simplification(exprf);
        exprg_old:= exprg;
        solan  := 0;
    Step 2: Combine simple simplification and simple expansion until get
    smallest expression or solan > 3.
        2.1 Expand simply exprg.
                exprh :=  Expand simply (exprg);
                If cannot expand then goto Step 3.
        2.2 Simplify simply exprh
                exprl:= Simplify simply(exprh);
                If length(exprl) < length(exprg_old) then
                        exprg_old := exprl;
                        solan := 0;
                Else,
                        solan := solan +1;
                end: # 2.2
        exprg := exprl;
        Go to step 2.
    Step 3: exprg := exprg_old; return exprg;
```

Here are simple simplification algorithm and simple expansion algorithm. The idea of both two algorithms is to perform the inference process combines the heuristic rules to enhance the speed of solving problems and achieve a better solution.

Simple Simplification Algorithm

```
Input:  exprf
Output: exprg
    Step 1: exprg  := exprf;
    Step 2: Using heuristic rules to select a rule r in Rrg that can be applied
    to exprg
        while ( r is found ) do
            exprg := r(exprg); // r(exprg) is a result after applying r to exprg
            record r;
            Using heuristic rules to select a rule r in Rrg that can be applied
on exprg.
        od;
    Step 3: Factor exprg.
            If can factor then goto step 2.
    Step 4: return exprg;
```

Some heuristic rules were used for selecting simplifying rule are listed below:

- Chose rule whose right – hand side is constant.
- Chose rule that has no new function in it
- Chose rule that involves functions in expression.
- Chose rule that length of its left – hand side is longest and length of its right – hand side is shortest.

Simple Expansion Algorithm

```
Input: exprf
Output: exprg
    Step 1: exprg  := exprf;
    Step 2: Using heuristic rules to select a rule r in Rkt that can be applied
    on exprg
            If r is found then exprg := r(exprg);  return exprg;
    Step 3: Transform product into sum in exprg
            If can transform then return exprg after transforming;
    Step 4: Provide the common denominator in exprg
            If can provide then return exprg after providing
    Step 5: return exprg;
```

Some heuristic rules were used for selecting expansion rule are listed below:

- Chose rule that has no new function in it.
- Chose rule that involves functions in expression.
- Chose rule makes expression has shortest length after applying.

4 Application

Based on the model of functional knowledge FKM, we built an automatic simplification program of trigonometric expressions. This program was implemented by language Maple and C #. It can reason automatically based on rules in order to simplify expression. In addition, it can present the solution in step by step, explaining the process of simplification. Moreover, the program has been tested on many exercises in trigonometry in math books in high education.

Here are some examples of simplification of trigonometric expression.

Example3: Simplify $f = \sqrt{\sin(x)^2(1+\cot(x)) + c\mathrm{os}(x)^2(1+\tan(x))}$

with $\left\{0 < x, x < \frac{\pi}{2}\right\}$

Solution:

Step 1: Apply $\tan(x) = \frac{\sin(x)}{c\mathrm{os}(x)}$

Then: $f = \sqrt{\sin(x)^2(1+\cot(x)) + c\mathrm{os}(x)^2\left(1+\frac{\sin(x)}{c\mathrm{os}(x)}\right)}$

Step 2: Apply $\cot(x) = \frac{c\mathrm{os}(x)}{\sin(x)}$

Then: $f = \sqrt{\sin(x)^2\left(1+\frac{\cos(x)}{\sin(x)}\right) + c\mathrm{os}(x)^2\left(1+\frac{\sin(x)}{c\mathrm{os}(x)}\right)}$

Step 3: Product to sum

Then: $f = \sqrt{c\mathrm{os}(x)^2 + 2\cos(x)\sin(x) + \sin(x)^2}$

Step 4: Apply $c\mathrm{os}(x)^2 + 2\cos(x)\sin(x) + \sin(x)^2 = \left(c\mathrm{os}(x) + \sin(x)\right)^2$

Then: $f = \sqrt{\left(c\mathrm{os}(x) + \sin(x)\right)^2}$

Step 5: Apply $\left\{0 < x, x < \frac{\pi}{2}\right\}$

Then: $f = c\mathrm{os}(x) + \sin(x)$

Compare with Maple

Maple is one of current excellent computer algebra systems that support automatic simplification of trigonometric expressions. However, its ability is limited in complex exercises. Moreover, Maple cannot provide a step by step solution. Therefore, in this section we do a comparison between Maple and the simplification program that built based on functional knowledge model FKM. Our program is implemented by Maple and C #. The following examples show that Maple cannot provide a step by step solution, or give a bad result, or cannot simplify in some cases. Then, it finds that the ability of our program is better than Maple's. Its solutions are in step by step and close to human thought.

Example 4: Simplify $f=\left(\tan(x)+\cot(x)\right)^2-\left(\tan(x)-\cot(x)\right)^2$

Result of Maple: 4

Result of program: 4

Solution:

Step 1: Apply $\left(\tan(x)+\cot(x)\right)^2=\cot(x)^2+2\tan(x)\cot(x)+\tan(x)^2$

Then: $f=\cot(x)^2+2\tan(x)\cot(x)+\tan(x)^2-\left(\tan(x)-\cot(x)\right)^2$

Step 2: Apply $\left(\tan(x)-\cot(x)\right)^2=\cot(x)^2-2\tan(x)\cot(x)+\tan(x)^2$

Then: $f=4\tan(x)\cot(x)$

Step 3: Apply $\tan(x)=\dfrac{\sin(x)}{\cos(x)}$

Then: $f=\dfrac{4\sin(x)\cot(x)}{\cos(x)}$

Step 4: Apply $\cot(x)=\dfrac{\cos(x)}{\sin(x)}$

Then: $f=4$

Example 5: Simplify $f=\sqrt{\sin(x)^2(1+\cot(x))+\cos(x)^2(1+\tan(x))}$

with $\left\{0<x,x<\dfrac{\pi}{2}\right\}$

Result of Maple: $\sqrt{2\cos(x)\sin(x)+1}$

Result of program: $f=\cos(x)+\sin(x)$

Solution: (see example 3)

Example 6: Simplify $f=\dfrac{\cos(x)^2-\sin(y)^2}{\sin(x)^2\sin(y)^2}-\cot(x)^2\cot(y)^2$

Result of Maple: $\dfrac{-\cos(x)^2+1-\cos(y)^2+\cos(x)^2\cos(y)^2}{\sin(x)^2\sin(y)^2}$

Result of Maple: -1

Solution:

Step 1: Apply $\cot(x)=\dfrac{\cos(x)}{\sin(x)}$

Then: $f=\dfrac{\cos(x)^2-\sin(y)^2}{\sin(x)^2\sin(y)^2}-\dfrac{\cos(x)^2\cot(y)^2}{\sin(x)^2}$

Step 2: Apply $\cot(y) = \frac{\cos(y)}{\sin(y)}$

Then: $$f = \frac{\cos(x)^2 - \sin(y)^2}{\sin(x)^2 \sin(y)^2} - \frac{\cos(x)^2 \cos(y)^2}{\sin(x)^2 \sin(y)^2}$$

Step 3: Product to sum

Then: $$f = \frac{\cos(x)^2}{\sin(x)^2 \sin(y)^2} - \frac{1}{\sin(x)^2} - \frac{\cos(x)^2 \cos(y)^2}{\sin(x)^2 \sin(y)^2}$$

Step 4: Common factor

Then: $$f = -\frac{\cos(x)^2 \left(-1 + \cos(y)^2\right)}{\sin(x)^2 \sin(y)^2} - \frac{1}{\sin(x)^2}$$

Step 5: Apply $-1 + \cos(y)^2 = -\sin(y)^2$

Then: $$f = \frac{\cos(x)^2}{\sin(x)^2} - \frac{1}{\sin(x)^2}$$

Step 6: Common factor

Then: $$f = \frac{-1 + \cos(x)^2}{\sin(x)^2}$$

Step 7: Apply $-1 + \cos(x)^2 = -\sin(x)^2$

Then: $f = -1$

5 Conclusion

Functional knowledge model is a new method for representing knowledge domains related to the functions, and their computing relations. This model is suitable for designing a knowledge base of functions in many different applications. With the explicit structure, model can be used to build the solving modules. This model has contributed to the development of methods for representing human knowledge on computers.

The technique of simplification above is based on the length of the expression. It has helped to solve the problem of simplifying automatically expression in many practical knowledge domains with many forms of expressions.

The program of automatic trigonometric simplifying has been tested on many exercises in trigonometry in math books of high education. This program can simplify many forms of expressions and provide solutions that are natural, accurate, and similar to the thinking of people.

References

1. Sowa, J.F.: Knowledge Representation - Logical, Philosophical, and Computational Foundations. Brooks/Cole, California (2000)
2. Tim Jones, M.: Aritificial Intelligence - A Systems Approach. Infinity Science Press LLC (2008)

3. Tyugu, E.: Algorithms and Architectures of Aritificial Intelligence. IOS Press (2007)
4. Lakemeyer, G., Nebel, B.: Foundations of Knowledge representation and Reasoning. Springer, Heidelberg (1994)
5. van Harmelen, F., Lifschitz, V., Porter, B. (eds.): Handbook of Knowledge Representation. Elsevier (2008)
6. Calmet, J., Tjandra, I.A.: Representation of Mathematical Knowledge. In: Raś, Z.W., Zemankova, M. (eds.) ISMIS 1991. LNCS, vol. 542, pp. 469–478. Springer, Heidelberg (1991)
7. Brewster, C., O'Hara, K.: Knowledge Representation with Ontologies: The Present and Future. IEEE Intelligent Systems 19(1), 72–73 (2004)
8. Sabine, B., Gills, K., Gills, M.: An Ontological approach to the construction of problem-solving models, Laria Research Report: LRR 2005-03 (2005)
9. Van Nhon, D.: A system that supports studying knowledge and solving of analytic geometry problems. In: 16th World Computer Congress, Proceedings of Conference on Education Uses of Information and Communication Technologies, Beijing, China, pp. 236–239 (2000)
10. Do, N.: An Ontology for Knowledge Representation and Applications. Proceeding of World Academy of Science, Engineering and Technology 32 (2008)
11. Fu, H., Zhong, X., Zeng, Z.: Automated and readable simplification of trigonometric expressions. Mathematical and Computer Modeling 44, 1169–1177 (2006)

Local Neighbor Enrichment for Ontology Integration

Trong Hai Duong[1], Hai Bang Truong[2], and Ngoc Thanh Nguyen[3]

[1] Faculty of Mathematics and Informatics,
Quangbinh University, Vietnam
haiduongtrong@gmail.com

[2] University of Information Technology,
VNU-HCM, Vietnam
bangth@uit.edu.vn

[3] Institute of Informatics,
Wroclaw University of Technology, Poland
Ngoc-Thanh.Nguyen@pwr.edu.pl

Abstract. The main aim of this research is to deal with enriching conceptual semantic by expanding local conceptual neighbor. The approach consists of two phases: neighbor enrichment phase and matching phase. The enrichment phase is based on analysis of the extension semantic the ontologies have. The extension we make use of in this work is generated an contextually expanded neighbor of each concept from external knowledge sources such as WordNet, ODP, and Wikimedia. Outputs of the enrichment phase are two sets of contextually expanded neighbors belonging to these two corresponding ontologies, respectively. The matching phase calculates similarities between these contextually expended neighbors, which yields decisions which concepts are to be matched.

Keywords: knowledge integration, information system, ontology integration.

1 Introduction

In the literature, there are many definitions of ontology integration, but the definition given by [16] is refereed in this research that defined as *the process of finding commonalities between two different ontologies O and O' and deriving a new ontology O^* that facilitates interoperability between computer systems that are base on the O and O' ontologies. The new ontology O^* may replace O or O', or it may be used only as an intermediary between a system based on O and system based on O'*. Finding ontological commonality is a very complex task, since ontologies have varies characteristics, e.g., the same concept but different names, the same name but different concepts, overlapping concepts but different concepts, multiple forms of the same concept, and multiple concepts of the same form [6]. Although several efforts in ontology integration have already been contributed, they have different focuses, assumptions, and limitations. According

J.-S. Pan, S.-M. Chen, N.T. Nguyen (Eds.): ACIIDS 2012, Part I, LNAI 7196, pp. 156–166, 2012.

to the literature, ontological similarity techniques that have been explored for commonalities finding can be classified into following four methods: instance-based similarity that is a similarity measurement between concepts based on concepts' common instances [4], lexical-based similarity that the similarity between two concepts is determined by analyzing the linguistic meanings of associated names [8,9], schema-based similarity that a similarity between two concepts is determined by analyzing the similarity between associated properties [1], and taxonomy-based similarity that is found by analyzing the structural relationships between them, such as subsumption [7].

The main aim of this research is to deal with enriching conceptual semantic by expanding local conceptual neighbor. The approach consists of two phases: neighbor enrichment phase and matching phase. The enrichment phase is based on analysis of the extension semantic the ontologies have. The extension we make use of in this work is generated an contextually expanded neighbor of each concept from external knowledge sources such as WordNet[1], ODP[2], and Wikimedia[3]. The intuition is that given two ontologies which should be matched, we construct a contextually expanded neighbor for each concept. The neighbor, is first derived from external knowledge sources, as it reflects the domain-independent concept for the corresponding concept, called domain-independent concept neighbor or global neighbor. Then, this independent neighbor matches to local neighbor (extracted from the corresponding ontology) to generate the corresponding contextually expanded neighbor. Outputs of the enrichment phase are two sets of contextually expanded neighbors, which belong to the two corresponding ontologies, respectively. The matching phase calculates similarities between these contextually expended neighbors, which yields decisions which concepts are to be matched.

2 Related Works

To enrich semantic for ontology, most previous works focus on using external knowledge to generate a enrichment structure such as a feature vector [17] and a forest for representing each concept [15].

The main task of the work [17] for the semantic enrichment structure is to generate a feature vector of a concept, which comes out as the result of extension analysis of the relevant ontologies.

In particularly, each concept is considered as a query to collect relevant documents that is then assigned to the query concept. However, if documents have already been assigned to specific concepts as their instances, we can skip this step and construct feature vector for the concepts directly. There are two steps to construct feature vector for each concept as follows: First step, using the vector space model tf/idf to construct the generalization of the documents. Second step, for each leaf concept, the feature vector is calculated as the feature vector

[1] http://wordnet.princeton.edu/

[2] http://www.dmoz.org/

[3] http://en.wikipedia.org

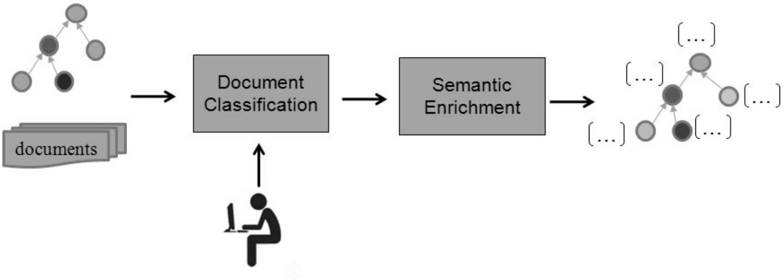


Fig. 1. Overview of the semantic enrichment process

of set of documents associated with the concept. For each none-leaf concept, the feature vector is calculated by taking into consideration contributions from the documents that have been assigned to the concept and its direct sub concepts. Similarities between concepts across different ontologies are determined by their feature vector similarities. The semantic enrichment process of this approach is shown in Fig. 1. This approach makes a strong assumption that the documents querying from a concept name are relevant to the concept. However, even the assumption is satisfied, the representative feature of a concept is common understanding or domain-independent the corresponding concept. Considering concepts in a specific context on a specific domain ontology, this approach will lead to mismatching as aforementioned.

Another system [15], which calls BLOOMS, is based on the idea of bootstrapping information on the LOD cloud. This approach is an utilization of the Wikipedia category hierarchy. Two ontologies, which are assumed to contain schema information, should be mapped by BLOOMS. It then proceeds with the following steps:

- Pre-processing: The extraction of concepts from the input ontologies are performed by removing property restrictions, individuals, and properties. Each concept name is tokenized to obtain a list of all simple words contained within them, with stop words removed.
- Construction: Each word belonging to a concept is a root of a tree constructing using information from Wikipedia. The depth of the tree limits to four based on empirical observation depths beyond four typically include very general categories, which are not useful for alignment. Therefore, each concept is represented by a forest including trees rooting by its tokenized words.
- Matching: The comparison of constructed BLOOMS forests is executed to determine which conceptual names are to be aligned.
- Post-processing: Enhancing results using the Alignment API and other reasoners.

Both aforementioned approaches used additional knowledge such as wikipedia or text corpus to enrich a concept semantic that yields to determine which concepts between two ontologies to align. However, these methods do not consider contextual concept, which leads to many mismatching concepts.

3 Methodologies

In matching problem, possible commonalities between two ontologies are determined by the similarity of ontological entity names leading to semantic mismatches and logical inconsistencies. The name of a concept may not express the precise semantics of the concept. Although several efforts in ontology matching have already been contributed to enrich semantic concepts for ontology integration such as [17,15]. However, the contextual concept has not yet considered leading to many mismatching concepts. In practice, to determine the similarity between concepts, it should be considered each concept in its corresponding domain-specific ontology. For this, this work aim at contextually expanding the concepts' neighbors which yields decisions whether concepts are to be matched.

To do so, we make five assumptions as follows:

(1) A local neighbor of a concept, which is a set of children, parents, grandchildren, grandparents, uncle, and nephews of the concept belonging to the corresponding ontology.
(2) A global neighbor of a concept, which is a set of children, parents, grandchildren, and nephews of the concept belonging to WordNet ontology or ODP or Wikipedia which are domain-independent ontologies or sources.
(3) Local neighbor of a concept can be used to identify domain-specific concept for the concept in a context of the corresponding ontology.
(4) External knowledge, such as WordNet, ODP, and Wikimedia, can be used to discover domain-independent concept for a concept.
(5) Global neighbor of a concept reflects the domain-independent concept for the concept, therefore global neighbor of a concept contains its local neighbor.

The contextually expanded neighbor of a concept can be generated by combining local neighbor and global neighbor of the concept.

3.1 Local Neighbor Enrichment Phase

The enrichment phase is based on analysis of the extension semantic the ontologies have. The extension we make use of in this work is generated an contextually expanded neighbor of each concept from external knowledge sources such as WordNet, ODP, and Wikimedia. The intuition is that given two ontologies which should be matched, we construct a contextually expanded neighbor for each concept. The neighbor, is first derived from external knowledge sources, as it reflects the domain-independent concept for the corresponding concept, called domain-independent neighbor or global neighbor. Then, this global neighbor matches to local neighbor (extracted from the corresponding ontology) to generate the corresponding contextually expanded neighbor. Outputs of the enrichment phase are two sets of contextually expanded neighbors, which belong to the two corresponding ontologies, respectively. The local neighbor enrichment architecture is as shown in Fig. 2.

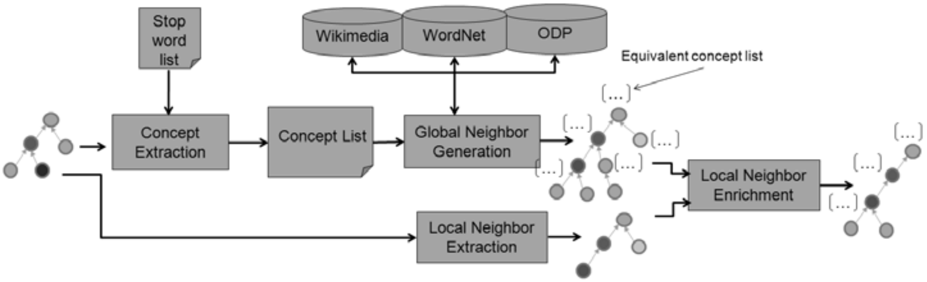


Fig. 2. Enrichment Architecture

Concept Extraction. Here, ontological concepts are considered as express names consisting of words in some natural language, e.g., English. A natural language processing technique is used to exploit morphological properties of each concept via its express name (called pre-process). A concept can be represented by a set of single words through the pre-process as shown in Fig. 3 [5].

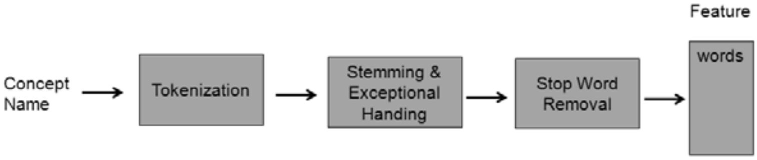


Fig. 3. Pre-Processing

- Tokenization: The concept names are segmented into sequences of tokens by a tokeniser which recognises punctuation, cases, blank characters, digits, etc. For example, *has_Hands_Free_Kits* becomes tokens{*has, Hands, Free, Kits*}.
- Stemming: The strings underlying tokens are morphologically analyzed in order to reduce them to normalised basic forms. For example, tokens $\{has, Ha-nds, Free, Kits\}$ becomes root form $\{have, hand, free, kit\}$.
- Stop word: Stop word is a word which is irrelevant to the content of the text, such as a pronoun, conjunction, preposition, article, and auxiliary verb. Removing stop words ensures more efficient processing by downsizing and the remaining words are relevant to the content. For example, after removing the stop words of the set $\{have, hand, free, kit\}$, it becomes $\{hand, free, kit\}$.

Global Neighbor Generation. To generate a global neighbor for a concept, we use several external knowledge sources as follows:

- The lexical database WordNet is particularly well suited to similarity measures, since it organizes nouns and verbs into hierarchies of is-a relations. In version 2.0, there are nine noun hierarchies comprising 80,000 concepts, and 554 verb hierarchies comprising 13,500 concepts [14].

- *ODP* is an open content directory of web pages maintained by a community of volunteer editors. It uses a taxonomic approach to represent topics and web pages that belong to these topics.
- Wikimedia service provides Wikipedia category hierarchy, which is a user-generated class hierarchy for Wikipedia pages. It also provides a search feature which we could generate a category hierarchy of a given concept.

A concept name can be considered as a noun phase. Therefore, it is easy to identify the head noun of a concept name [9]. We generate a longest concept name containing the head noun that occurs on either WordNet, ODP, or Wikimedia. Then, a global neighbor of the concept is derived from either WordNet, or ODP based on *ISA* relation (hierarchical relation), or Wikimedia by the way presented in [15].

Local Neighbor Extraction. To extract a local neighbor of a concept, we only collect a set of children, parents, grandchildren, grandparents, uncle, and nephews of the concept belonging to its corresponding ontology.

Local Neighbor Enrichment. Let us recall that the aforementioned previous approaches use a additional knowledge such as Wikipedia or text corpus to enrich the semantic of a concept, that yields to determine which classes between two ontologies to align. However, their representative structure for a concept covers the domain-independent concept. They do not consider the context that the concept belongs to a domain-specific ontology. Therefore, it leads to many mismatching concepts.

Here, we assume that the local neighbor of a concept can be used to identify specific domain for the concept in a context of the corresponding ontology. We align between the local neighbor and the global neighbor of the concept to identify a partial global neighbor in a context of the domain-specific ontology to which the concept belongs. We consider the partial global neighbor as the contextually expanded neighbor of the concept. Therefore, each concept is represented by a contextually expanded neighbor, which yields decisions which concepts are to be matched.

3.2 Matching Phase

The matching phase is to calculate similarities between these two sets of contextually expanded neighbors, which yields decisions which concepts are to be matched. However, we do not exhaustively compute similarities between the concepts belonging to different ontologies, a method to skip the unmatchable pairs of concepts by propagating priorly matchable concepts (*PMC*) is applied [10]. *PMC* is a collection of pairs of concepts belonging to two different ontologies in the same *Concept Type*, which are arranged in descending order according to the *Concept Importance* score similarity of pair of concepts. The role of *Concept Types* is to analyse the relations and constraints among concepts to classify concepts belonging to an ontology into disjoint categories. Two concepts in the same category cross different ontologies should be priorly checked similarity. It is useful to avoid checking unmatchable concepts and to solve mismatching on ontology integration. The *Concept Importance* is an importance measurement of a concept that shows how a central and richly described concept is in the hierarchy of the ontology. Therefore, two concepts across the ontologies should be priorly checked for similarity if their importance score similarity is greatest.

4 Experiments

4.1 Data Sets

We performed a comprehensive evaluation of the proposed system using third party datasets and other state-of-the-art systems in ontology matching. More specifically, we evaluated our proposed method in two different ways. Firstly, we examined the ability of the proposed method to serve as a general purpose ontology matching system, by comparing it with other systems on the Ontology Alignment Evaluation Initiative (OAEI)[4] benchmarks[5]. The domain of benchmarks test is Bibliographic references. The test is based on one particular ontology dedicated to the very narrow domain of bibliography and alterative ontologies of the same domain. They are organized in three groups: Data test 1 is consisting of simple tests (101-104), Data test 2 containing systematic tests (201-266), and Data test 3 including four real-life ontologies of bibliographic references (301-304) found on the web.

Secondly, we evaluated the proposed method for the purpose of LOD schema integration and compared it with other systems for ontology matching on LOD schema alignment. This data set contains schema-level mappings from two LOD ontologies to Proton (an upper level ontology) created manually by human experts for a real world application called FactForge, with over 300 classes and 100 properties. These two LOD ontologies include:

- DBpedia[6]: The RDF version of Wikipedia, created manually from Wikipedia article infoboxes. DBpedia consists of 259 classes ranging from general classes (e.g. Event) to domain specific ones (e.g. Protein).
- Geonames[7]: A geographic data set with over 6 million locations of interest, which are classified into 11 different classes.

4.2 Evaluation Method

Here, we use *Precision* and *Recall* measures to evaluate our approach in which N_{total} denotes the total number of pairs for matching concepts between the candidate ontologies by experts, $N_{correct}$ and $N_{incorrect}$ correspond to the number of correct pairs for matching concepts and the number of incorrect pairs for matching concepts sought by our system, respectively.

[4] http://oaei.ontologymatching.org

[5] http://oaei.ontologymatching.org/2010/benchmarks/

[6] http://downloads.dbpedia.org/3.5.1/dbpedia3.5.1.owl.bz2

[7] http://geonames.org

Precision is used to evaluate the ratio of incorrectly extracted relationships.

$$\frac{N_{correct}}{N_{correct} + N_{incorrect}} \tag{1}$$

Recall is used to evaluate the ratio of correct matching to the sought total by the system.

$$\frac{N_{correct}}{N_{total}} \tag{2}$$

Finally, the F-measure is calculated by combining precision and recall as follows:

$$F = 2 * \frac{precision * recall}{precision + recall} \tag{3}$$

4.3 Evaluation Result

In order to test the quality of matches generated using our proposed method in terms of two widely used matrices, Recall, and Precision. The proposed method compared with existing solutions that performed well for LOD ontology alignment [15]. These solutions include:

- BLOOMS: This method is based on the idea of bootstrapping information already present on the LOD cloud. BLOOMS uses a rich knowledge Wikipedia to enhance semantical concepts before matching decision [15].
- S-Match: This method utilizes three matching algorithms basic, minimal, and structure preserving to establish mappings between the classes of two ontologies [11].
- AROMA: This method utilizes the association rule mining paradigm to discover equivalence and subclass relationships between the classes of two ontologies [3].

Fig. 4 shows the results for all our method *Aprior* and the previous works *AROMA*, *BLOOMS*, and *S-Match*. Results show that our method performed well among the other methods on both precision and recall. Specially, the proposed method has extremely high precision. The reason is that the proposed algorithm avoids computing similarities between mismatching concepts based on *Concept Types* and *Concept Importance* measurement in [10]. However, no method performed well on aligning Geonames with Proton. The only matching found by our method (and the other methods) is the concept *Country* in Geonames is equivalent to the class *Nation* in Proton. The key reasons for the poor performance include: 1) Geonames has a small number of classes (and hence very limited contextual information) and 2) the names of the classes in Geonames are often vague and ambiguous, which made it difficult to compute their similarity.

We also ran our system on OAEI Brenchmark data set and compared its performance with the a representative method [15] for Benchmark data set (see Fig.5.). Our method performs well for this data set in comparison with above mentioned method. That means our method well performs in not only LOD purpose but also general purpose test such as OAEI Brenchmark data set.

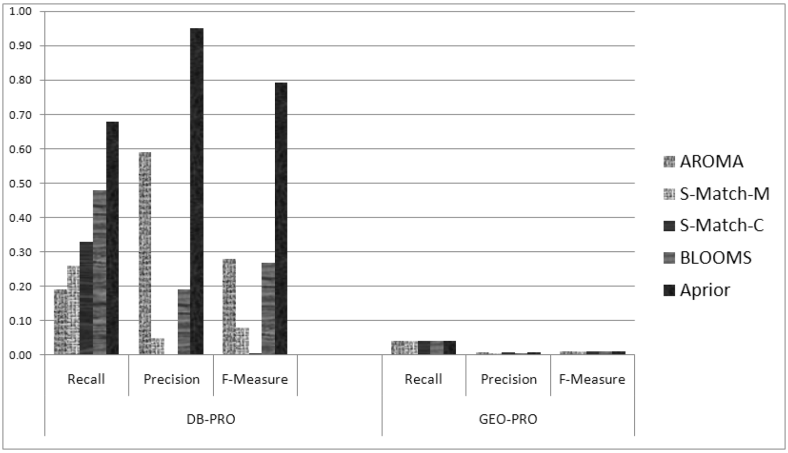


Fig. 4. LOD Schema Evaluation Comparison between our method and the previous works *BLOOMS, AROMA*, and *S-Match*

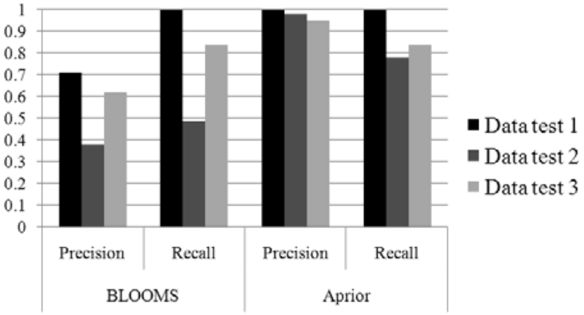


Fig. 5. Evaluation Comparison between our method *Aprior* and *BLOOMS* on Brenchmark data set

5 Conclusions

The main aim of this research is to deal with enriching conceptual semantic by expanding local conceptual neighbor. The extension we make use of in this work is generated an contextually expanded neighbor of each concept from external knowledge sources such as WordNet, ODP, and Wikimedia. Experimental results shows that our algorithm perform significantly in term of accuracy compare with some known methods from OAEI Benchmarks and Linked Open Data sets and quite better results compare with the existing methods.

Acknowledgement. This research was partially funded by Vietnam National Foundation for Science and Technology Development (NAFOSTED) under grant number 102.01- 2011.10.

This research was partially supported by Polish Ministry of Science and Higher Education under grant no. N N519 407437 (2009-2012).

References

1. Castano, S., Ferrara, A., Montanelli, S.: Matching Ontologies in Open Networked Systems: Techniques and Applications. In: Spaccapietra, S., Atzeni, P., Chu, W.W., Catarci, T., Sycara, K. (eds.) Journal on Data Semantics V. LNCS, vol. 3870, pp. 25–63. Springer, Heidelberg (2006)
2. Danilowicz, C., Nguyen, N.T.: Consensus-based methods for restoring consistency of replicated data. In: Kopotek, M., et al. (eds.) Advances in Soft Computing, Proceedings of 9th International Conference on Intelligent Information Systems 2000, pp. 325–336. Physica-Verlag (2000)
3. David, J., Guillet, F., Briand, H.: Matching directories and OWL ontologies with AROMA. In: CIKM 2006: The 15th ACM International Conference on Information and Knowledge Management, pp. 830–831. ACM, New York (2006)
4. Doan, A.H., Madhavan, J., Domingos, P., Halevy, A.: Ontollogy matching: a machine learning approach. In: Handbook on Ontologies in Information Systems, pp. 397–416. Springer, Heidelberg (2003)
5. Duong, T.H., Nguyen, N.T., Jo, G.S.: Effective Backbone Techniques for Ontology Integration. In: Nguyen, N.T., Szczerbicki, E. (eds.) Intelligent Systems for Knowledge Management. SCI, vol. 252, pp. 197–227. Springer, Heidelberg (2009)
6. Duong, T.H., Jo, G.S., Jung, J.J., Nguyen, N.T.: Complexity Analysis of Ontology Integration Methodologies: A Comparative Study. Journal of Universal Computer Science 15(4), 877–897 (2009)
7. Duong, T.H., Nguyen, N.T., Jo, G.S.: A Hybrid Method for Integrating Multiple Ontologies. Cybernetics and Systems 40(2), 123–145 (2009)
8. Duong, T.H., Nguyen, N.T., Jo, G.S.: A Method for Integration across Text Corpus and WordNet-based Ontologies. In: IEEE/ACM/WI/IAT 2008 Workshops Proceedings, pp. 1–4. IEEE Computer Society (2008)
9. Duong, T.H., Nguyen, N.T., Jo, G.S.: A Method for Integration of WordNet-based Ontologies Using Distance Measures. In: Lovrek, I., Howlett, R.J., Jain, L.C. (eds.) KES 2008, Part I. LNCS (LNAI), vol. 5177, pp. 210–219. Springer, Heidelberg (2008)
10. Duong, T.H., Jo, G.S.: Anchor-Prior: An Effective Algorithm for Ontology Integration. In: IEEE International Conference on Systems, Man, and Cybernetics (IEEE SMC 2011), pp. 942–947. IEEE Computer Society, Anchorage (2011)
11. 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.) ESWS 2004. LNCS, vol. 3053, pp. 61–75. Springer, Heidelberg (2004)
12. Maedche, A., Motik, B., Silva, N., Volz, R.: MAFRA – A MApping FRAmework for Distributed Ontologies. In: Gómez-Pérez, A., Benjamins, V.R. (eds.) EKAW 2002. LNCS (LNAI), vol. 2473, pp. 235–250. Springer, Heidelberg (2002)
13. Nguyen, N.T.: Conflicts of Ontologies – Classification and Consensus-Based Methods for Resolving. In: Gabrys, B., Howlett, R.J., Jain, L.C. (eds.) KES 2006, Part II. LNCS (LNAI), vol. 4252, pp. 267–274. Springer, Heidelberg (2006)
14. Pedersen, T., Patwardhan, S., Michelizzi, J.: WordNet::Similarity-measuring the relatedness of concepts. In: Proceedings of NAACL (2004)

15. Jain, P., Hitzler, P., Sheth, A.P., Verma, K., Yeh, P.Z.: Ontology Alignment for Linked Open Data. In: Patel-Schneider, P.F., Pan, Y., Hitzler, P., Mika, P., Zhang, L., Pan, J.Z., Horrocks, I., Glimm, B. (eds.) ISWC 2010, Part I. LNCS, vol. 6496, pp. 402–417. Springer, Heidelberg (2010)
16. Sowa, J.F.: Knowledge Representation: Logical, Philosophical and Computational Foundations. Brooks/Cole (2000)
17. 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)

A Novel Choquet Integral Composition Forecasting Model Based on M-Density

Hsiang-Chuan Liu[1,5], Shang-Ling Ou[2], Hsien-Chang Tsai[3],
Yih-Chang Ou[4], and Yen-Kuei Yu[5]

[1] Department of Biomedical Informatics, Asia University,
500, Lioufeng Rd. Wufeng, Taichung 41354, Taiwan, R.O.C.
[2] Department of Agronomy, National Chung Hsing University,
250, Kuo Kuang Rd., Taichung 40227, Taiwan, R.O.C.
[3] Department of Biology, National Changhua University of Education,
1, Jin-De Rd., Changhua City, 50007, Taiwan, R.O.C.
[4] Department of Finance, Ling Tung University,
1, Ling Tung Rd., Taichung, 40852, Taiwan, R.O.C.
[5] Graduate Institute of Educational Measurement and Statistics,
National Taichung University of Education,
140, Min-Shen Rd., Taichung 40306, Taiwan, R.O.C.
lhc@asia.edu.tw, slou@dragon.nchu.edu.tw,
bihft@cc.ncue.edu.tw, ycou@mail.ltu.edu.tw,
haikuei@gmail.com

Abstract. In this paper, a novel density, *M*-density, was proposed. Based on this new density, a novel composition forecasting model was also proposed. For comparing the forecasting efficiency of this new density with the well-known density, *N*-density, a real data experiment was conducted. The performances of Choquet integral composition forecasting model with extensional L-measure, λ-measure and P-measure, by using *M*-density and *N*-density, respectively, a ridge regression composition forecasting model and a multiple linear regression composition forecasting model and the traditional linear weighted composition forecasting model were compared. Experimental result showed that the Choquet integral composition forecasting model with respect to extensional L-measure based on *M*-density outperforms other composition forecasting models. Furthermore, for each fuzzy measure, including the L_E-measure, L-measure, λ-measure and P-measure, the *M*-density based Choquet integral composition forecasting model is better than the *N*-density based.

Keywords: Choquet integral, composition forecasting model, *M*-density, *N*-density, *R*-density.

1 Introduction

The composition forecasting model is first considered by the work of Bates and Granger (1969) [1], they are now in widespread use in many fields, especially in economic field. Zhang Wang and Gao (2008) [2] applied the linear composition

J.-S. Pan, S.-M. Chen, N.T. Nguyen (Eds.): ACIIDS 2012, Part I, LNAI 7196, pp. 167–176, 2012.

forecasting model which composed the time series model, the second-order exponential smoothing model and GM(1,1) forecasting model in the Agricultural Economy Research. In our previous work [6], we extended the work of Zhang, Wang, and Gao to propose some nonlinear composition forecasting model which also composed the time series model, the second-order exponential smoothing model and GM(1,1) forecasting model by using the ridge regression model [3] and the theory of Choquet integral with respect to some fuzzy measures, including extensional L-measure, L-measure, λ-measure and P-measure [4-10], and then found that the extensional L-measure Choquet integral based composition forecasting model is the best one. However, all of above mentioned Choquet integral composition forecasting models with some different fuzzy measures are based on *N*-density, we know that the performance of any Choquet integral is predominate by its fuzzy measure, and the performance of any fuzzy measure is predominate by its fuzzy density function, in other words, the performance of any Choquet integral is predominate by its fuzzy density function.

In this paper, a novel fuzzy density function, called *M*-density, is considered. Based on this new fuzzy density function, a novel composition forecasting model is also considered. For comparing the forecasting efficiency of this new fuzzy density function with the well-known fuzzy density functions, *N*-density, is also considered.

2 The Composition Forecasting Model

In this paper, for evaluating the forecasting validation of forecasting model to sequential data, the sequential mean square error is used, its formal definition is listed as follows.

Definition 1. Sequential Mean Square Error (SMSE) [6]

If θ_{t+j} is the realized value of target variable at time $(t+j)$, $\hat{\theta}_{t+j|t}$ is the forecasted value of target variable at time $(t+j)$ based on training data set from time 1 to time *t*,

and
$$SMSE\left(\hat{\theta}_t^{(h)}\right)=\frac{1}{h}\sum_{j=1}^{h}\left(\hat{\theta}_{t+j|t+j-1}-\hat{\theta}_{t+j}\right)^2 \tag{1}$$

then $SMSE\left(\hat{\theta}_t^{(h)}\right)$ is called the sequential mean square error (SMSE) of the *h* forecasted values of target variable from time $(t+1)$ to time $(t+h)$ based on training data set from time 1 to time *t*. The composition forecasting model or combination forecasting model can be defined as follows.

Definition 2. Composition Forecasting Model [1, 2, 6]

(i) Let y_t be the realized value of target variable at time *t*.

(ii) Let $x_{t,1}, x_{t,2}, ..., x_{t,m}$ be a set of *m* competing predictors of y_t, $\hat{y}_t$ be a function *f* of $x_{t,1}, x_{t,2}, ..., x_{t,m}$ with some parameters, denoted as

$$\hat{y}_t = f\left(x_{t,1}, x_{t,2}, ..., x_{t,m}\right) \tag{2}$$

(iii) Let $x_{t+j|t,k}$ be the forecasted values of y_t by competing predictor k at time $(t+j)$ based on training data set from time 1 to time *t*, and for the same function *f* as above,

Let
$$\hat{y}_{t+j|t} = f\left(x_{t+j,1}, x_{t+j,2}, ..., x_{t+j,m}\right) \tag{3}$$

(iv) Let
$$SMSE\left(\hat{y}_t^{(h)}\right) = \frac{1}{h}\sum_{j=1}^{h}\left(\hat{y}_{t+j|t+j-1} - y_{t+j}\right)^2 \tag{4}$$

$$SMSE\left(x_{t,k}^{(h)}\right) = \frac{1}{h}\sum_{j=1}^{h}\left(x_{t+j,k} - y_{t+j}\right)^2 \tag{5}$$

For current time *t* and the future *h* times, if

$$SMSE\left(\hat{y}_t^{(h)}\right) \leq \min_{1\leq k\leq m} SMSE\left(x_{t,k}^{(h)}\right) \tag{6}$$

then $\hat{y}_t$ is called a composition forecasting model for the future *h* times of $x_{t,1}, x_{t,2}, ..., x_{t,m}$ or, in brief, a composition forecasting model of $x_{t,1}, x_{t,2}, ..., x_{t,m}$.

Definition 3. Linear Combination Forecasting Model [1, 6]

For given parameters $\beta_k \in R, \sum_{k=1}^{m}\beta_k = 1$, let

$$\hat{y}_t = \sum_{k=1}^{m}\beta_k x_{t,k} \tag{7}$$

If $\hat{y}_t$ is a composite forecasting model of $x_{t,1}, x_{t,2}, ..., x_{t,m}$ then $\hat{y}_t$ is called a linear combination forecasting model or linear composition forecasting model, otherwise, it is called a non-linear combination forecasting model or non-linear composition forecasting model.

Definition 4. Ridge Regression Composition Forecasting Model [3, 6]

(i) Let $\underline{y}_t = \left(y_1, y_2, ..., y_t\right)^T$ be realized data vector of target variable from time 1 to time *t*, $\underline{x}_{t,k} = \left(x_{1,k}, x_{2,k}, ..., x_{t,k}\right)^T$ be a forecasted value vector of competing predictor *k* of target variable y_t from time 1 to time *t*.

(ii) Let X_t be a forecasted value matrix of m competing predictors of target variable y_t from time 1 to time t.

(iii) Let
$$\underline{\hat{y}}_t = \left(\hat{y}_1, \hat{y}_2, ..., \hat{y}_t\right)^T \tag{8}$$

$$f\left(X_t\right) = f\left(\underline{x}_{t,1}, \underline{x}_{t,2}, ..., \underline{x}_{t,m}\right) \tag{9}$$

(iv) Let
$$\underline{\beta}_t^{(r)} = \left(\beta_{t,1}^{(r)}, \beta_{t,2}^{(r)}, ..., \beta_{t,m}^{(r)}\right)^T = \left(X_t^T X_t + rI_m\right)^{-1} X_t^T \underline{y}_t \tag{10}$$

$$\underline{\hat{y}}_t = f\left(X_t\right) = X_t \underline{\beta}_t^{(r)} \tag{11}$$

Then
$$\underline{\hat{y}}_{t+j|t} = f\left(X_{t+j}\right) = X_{t+j} \underline{\beta}_t^{(r)} \tag{12}$$

$$\begin{aligned} \hat{y}_{t+j|t} &= f\left(x_{t+j,1}, x_{t+j,2}, ..., x_{t+j,m}\right) \\ &= \left[x_{t+j,1}, x_{t+j,2}, ..., x_{t+j,m}\right] \underline{\beta}_t^{(r)} = \sum_{k=1}^{m} \beta_{t,k}^{(r)} x_{t+j,k} \end{aligned} \tag{13}$$

For current time t and the future h times, if

$$SMSE\left(\hat{y}_t^{(h)}\right) \le \min_{1 \le k \le m} SMSE\left(x_{t,k}^{(h)}\right) \tag{14}$$

And ridge coefficient $r = 0$ then $\hat{y}_t$ is called a multiple linear regression combination forecasting model of $x_{t,1}, x_{t,2}, ..., x_{t,m}$. If formula (14) is satisfied and $r > 0$, then $\hat{y}_t$ is called a ridge regression composition forecasting model of $x_{t,1}, x_{t,2}, ..., x_{t,m}$. Note that Hoerl, Kenard, and Baldwin (1975) suggested that the ridge coefficient of ridge regression is

$$r = \frac{m\hat{\sigma}^2}{\underline{\beta}_t^T \underline{\beta}}, \quad \hat{\sigma}^2 = \frac{1}{t}\sum_{i=1}^{t}\left(y_i - \hat{y}_t\right)^2 \tag{15}$$

3 Choquet Integral Composition Forecasting Model

3.1 Fuzzy Measures

Definition 5. Fuzzy Measure [4-10]
A fuzzy measure μ on a finite set X is a set function $\mu : 2^X \rightarrow [0,1]$ satisfying the following axioms:

$$\mu(\phi)=0,\mu(X)=1 \qquad \text{(boundary conditions)} \tag{16}$$

$$A\subseteq B\Rightarrow\mu(A)\le\mu(B) \qquad \text{(monotonicity)} \tag{17}$$

3.2 Fuzzy Density Function

Definition 6. Fuzzy Density Function, Density [4-6]

(i) A fuzzy density function of a fuzzy measure μ on a finite set X is a function $d: X\rightarrow[0,1]$ satisfying:

$$d(x)=\mu(\{x\}),x\in X \tag{18}$$

$d(x)$ is called the density of singleton x.

(ii) A fuzzy density function is called a normalized fuzzy density function or a density if it satisfying

$$\sum_{x\in X}d(x)=1 \tag{19}$$

Definition 7. Standard Fuzzy Measure [4-6]

A fuzzy measure is called a standard fuzzy measure, if its fuzzy density function is a normalized fuzzy density function.

Definition 8. *N*-density

Let μ be a fuzzy measure on a finite set $X=\{x_1,x_2,...,x_n\}$, y_i be global response of subject *i* and $f_i(x_j)$ be the evaluation of subject *i* for singleton x_j, satisfying:

$$0<f_i(x_j)<1, i=1,2,...,N,\ \ j=1,2,...,n \tag{20}$$

If

$$d_N(x_j)=\frac{r(f(x_j))}{\sum_{j=1}^{n}r(f(x_j))}, j=1,2,...,n \tag{21}$$

Where $r(f(x_j))$ is the linear regression coefficient of y_i on $f(x_j)$,

then the function $d_N: X\rightarrow[0,1]$ satisfying $\mu(\{x\})=d_N(x),\forall x\in X$ is a fuzzy density function, called *N*-density of μ. Note that *N*-density is a normalized fuzzy density function.

3.3 *M*-density

In this paper, a novel normalized fuzzy density function based on Mean Square Error, denoted *M*-density, is proposed, its formal definition is introduced as follows:

Definition 9. ***M*-density**
Let μ be a fuzzy measure on a finite set $X = \{x_1, x_2, ..., x_n\}$, y_i be global response of subject *i* and $f_i(x_j)$ be the evaluation of subject *i* for singleton x_j, satisfying:

$$0 < f_i(x_j) < 1, i = 1,2,...,N, \; j = 1,2,...,n \tag{22}$$

If
$$d_M(x_j) = \frac{[MSE(x_j)]^{-1}}{\sum_{j=1}^{n}[MSE(x_j)]^{-1}}, j = 1,2,...,n \tag{23}$$

Where
$$MSE(x_j) = \frac{1}{N}\sum_{i=1}^{N}(y_i - f_i(x_j))^2 \tag{24}$$

then the function $d_M : X \to [0,1]$ satisfying $\mu(\{x\}) = d_M(x), \forall x \in X$ is a fuzzy density function, and called *M*-density of μ.

3.4 λ-measure

Definition 10. λ-measure [9]
For a given fuzzy density function d, a λ-measure, g_λ, is a fuzzy measure on a finite set *X*, satisfying:

$$\begin{aligned} &A, B \in 2^X, A \cap B = \phi, A \cup B \neq X \\ &\Rightarrow g_\lambda(A \cup B) = g_\lambda(A) + g_\lambda(B) + \lambda g_\lambda(A) g_\lambda(B) \end{aligned} \tag{25}$$

$$\prod_{i=1}^{n}[1 + \lambda d(x_i)] = \lambda + 1 > 0, \; d(x_i) = g_\lambda(\{x_i\}) \tag{26}$$

Theorem 1. For any given normalized fuzzy density function, a λ-measure is just an additive measure.

3.5 P-measure

Definition 11. P-measure [10]
For a given fuzzy density function *d*, a P-measure, g_P, is a fuzzy measure on a finite set *X*, satisfying:

$$^{\forall}A\in 2^{X}\Rightarrow g_P\left(A\right)=\max_{x\in A}d\left(x\right)=\max_{x\in A}g_P\left(\left\{x\right\}\right) \tag{27}$$

3.6 L-measure

Definition 12. L-measure [4]
For a given fuzzy density function *d*, L-measure, g_L, is a measure on a finite set *X*, $\left|X\right|=n$, satisfying:

$$L\in\left[0,\infty\right),A\subset X\Rightarrow g_L=\max_{x\in A}d\left(x\right)+\frac{\left(\left|A\right|-1\right)L\sum_{x\in A}d\left(x\right)\left[1-\max_{x\in A}d\left(x\right)\right]}{\left[n-\left|A\right|+L\left(\left|A\right|-1\right)\right]\sum_{x\in X}d\left(x\right)} \tag{28}$$

Theorem 2. Important Properties of L-measure [4]

(i) For any $\mathrm{L}\in\left[0,\infty\right)$, L-measure is a fuzzy measure, in other words, L-measure has infinite fuzzy measure solutions.
(ii) L-measure is an increasing function on L.
(iii) if $\mathrm{L}=0$ then L-measure is just the P-measure.
Note that L-measure contains additive measure and λ-measure.

3.7 Extensional L-measure

Definition 13. Extensional L-measure, L$_E$-measure [5]
For a given fuzzy density function *d*, an extensional L-measure, or L$_E$-measure, g_{LE}, is a measure on a finite set $X,\left|X\right|=n$, satisfying:

$$L\in\left[-1,\infty\right),A\subset X$$
$$\Rightarrow g_{LE}\left(A\right)=\begin{cases}\left(1+L\right)\sum_{x\in A}d\left(x\right)-L\max_{x\in A}d\left(x\right) & ,L\in\left[-1,0\right]\\ \sum_{x\in A}d\left(x\right)+\frac{\left(\left|A\right|-1\right)L\sum_{x\in A}d\left(x\right)\left[1-\sum_{x\in A}d\left(x\right)\right]}{\left[n-\left|A\right|+L\left(\left|A\right|-1\right)\right]\sum_{x\in X}d\left(x\right)} & ,L\in\left(0,\infty\right)\end{cases} \tag{29}$$

Theorem 3. Important Properties of L$_E$ –measure [5]

(i) For any $\mathrm{L}\in\left[-1,\infty\right)$, L$_E$ -measure is a fuzzy measure, in other words, L$_E$-measure has infinite fuzzy measure solutions.
(ii) L$_E$-measure is an increasing function on L.
(iii) if $\mathrm{L}=-1$ then L$_E$-measure is just the P-measure.
(iv) if $\mathrm{L}=0$ then L$_E$-measure is just the additive measure.

(v) if $\mathrm{L}=0$ and $\sum_{x\in X} d(x)=1$, then L$_E$-measure is just the λ-measure.

(vi) if $-1<\mathrm{L}<0$ then L$_E$-measure is a sub-additive measure.

Note that additive measure, λ-measure and P-measure are two special cases of L$_E$-measure.

3.8 Choquet Integral

Definition 14. Choquet Integral [7-8]

Let μ be a fuzzy measure on a finite set $X=\{x_1,x_2,...,x_m\}$. The Choquet integral of $f_i : X \to R_+$ with respect to μ for individual *i* is denoted by

$$\int_C f_i d\mu = \sum_{j=1}^{m}\left[f_i\left(x_{(j)}\right)-f_i\left(x_{(j-1)}\right)\right]\mu\left(A_{(j)}^i\right),\ i=1,2,...,N \tag{30}$$

where $f_i\left(x_{(0)}\right)=0$, $f_i\left(x_{(j)}\right)$ indicates that the indices have been permuted so that

$$0\le f_i\left(x_{(1)}\right)\le f_i\left(x_{(2)}\right)\le ...\le f_i\left(x_{(m)}\right),\ A_{(j)}=\left\{x_{(j)},x_{(j+1)},...,x_{(m)}\right\} \tag{31}$$

Theorem 4. If a λ-measure is a standard fuzzy measure on $X=\{x_1,x_2,...,x_m\}$, and $d : X \to [0,1]$ is its fuzzy density function, then the Choquet integral of $f_i : X \to R_+$ with respect to λ for individual *i* satisfying

$$\int_C f_i d\lambda = \sum_{j=1}^{m} d\left(x_j\right) f_i\left(x_j\right),\ i=1,2,...,N \tag{32}$$

3.9 Choquet Integral Composition Forecasting Model

Definition 15. Choquet Integral Composition Forecasting Model [6]

(i) Let y_t be the realized value of target variable at time *t*,

(ii) Let $X=\{x_1,x_2,...,x_m\}$ be the set of m competing predictors,

(iii) Let $f_t : X \to R_+$, $f_t(x_1), f_t(x_2),..., f_t(x_m)$ be *m* forecasting values of y_t by competing predictors $x_1,x_2,...,x_m$ at time *t*.

If μ is a fuzzy measure on X, $\alpha,\beta\in R$ satisfying

$$\left(\hat{\alpha},\hat{\beta}\right)=\arg\min_{\alpha,\beta}\left[\sum_{t=1}^{N}\left(y_i-\alpha-\beta\int_C f_t dg_\mu\right)\right] \tag{33}$$

$$\hat{\alpha} = \frac{1}{N}\sum_{t=1}^{N} y_t - \hat{\beta}\frac{1}{N}\sum_{t=1}^{N}\int f_t dg_\mu \,,\quad \hat{\beta} = \frac{S_{yf}}{S_{ff}} \tag{34}$$

then $\hat{y}_t = \hat{\alpha} + \hat{\beta}\int f_t dg_\mu, t = 1,2,...,N$ is called the Choquet integral regression composition forecasting estimator of y_t, and this model is also called the Choquet integral regression composition forecasting model with respect to μ -measure.

Theorem 5. If a λ-measure is a standard fuzzy measure then Choquet integral regression composition forecasting model with respect to λ-measure is just a linear combination forecasting model.

4 Experiments and Results

A real data of the grain production with 3 kinds of forecasted values of the time series model, the exponential smoothing model and GM(1,1) forecasting model, respectively, in Jilin during 1952 to 2007 was obtained from the Table 1. in the paper of Zhang, Wang and Gao [2]. For evaluating the proposed new density based composition forecasting model, an experiment with the above-mentioned data by using sequential mean square error was conducted.

We arrange the first 50 years grain production and their 3 kinds of forecasted values as the training set and the rest data as the forecasting set. And the following *N*-density and *M*-density of all fuzzy measures were used

$$N\text{-density:} \qquad \{0.3331, \quad 0.3343, \quad 0.3326\} \tag{35}$$

$$M\text{-density:} \qquad \{0.2770, \quad 0.3813, \quad 0.3417\} \tag{36}$$

The performances of Choquet integral composition forecasting model with extensional L-measure, L-measure, λ-measure and P-measure, respectively, a ridge regression composition forecasting model and a multiple linear regression composition forecasting model and the traditional linear weighted composition forecasting model were compared. The result is listed in Table 1.

Table 1. SMSEs of 2 densities for 6 composition forecasting models

Composition forecasting Models		SMSE	
		N-density	*M*-density
Choquet integral regression	L_E-measure	13939.84	13398.29
	L-measure	14147.83	13751.60
	λ-measure	21576.38	19831.86
	P-measure	16734.88	16465.98
Ridge regression		18041.92	
Multiple linear regression		24438.29	

Table 1 shows that the *M*-density based Choquet integral composition forecasting model with respect to L_E-measure outperforms other composition forecasting models. Furthermore, for each fuzzy measure, including the L_E-measure, L-measure, λ-measure and P-measure, the *M*-density based Choquet integral composition forecasting model is better than the *N*-density based.

5 Conclusion

In this paper, a new density, *M*-density, was proposed. Based on *M*-density, a novel composition forecasting model was also proposed. For comparing the forecasting efficiency of this new density with the well-known density, *N*-density, a real data experiment was conducted. The performances of Choquet integral composition forecasting model with extensional L-measure, λ-measure and P-measure, by using *M*-density and *N*-density, respectively, a ridge regression composition forecasting model and a multiple linear regression composition forecasting model and the traditional linear weighted composition forecasting model were compared. Experimental result showed that for each fuzzy measure, including the L_E-measure, L-measure, λ-measure and P-measure, the *M*-density based Choquet integral composition forecasting model is better than the *N*-density based, and the *M*-density based Choquet integral composition forecasting model outperforms all of other composition forecasting models.

Acknowledgment. This paper is partially supported by the grant of National Science Council of Taiwan Government (NSC 100-2511-S-468-001).

References

1. Bates, J.M., Granger, C.W.J.: The Combination of Forecasts. Operations Research Quarterly 4, 451–468 (1969)
2. Zhang, H.-Q., Wang, B., Gao, L.-B.: Application of Composition Forecasting Model in the Agricultural Economy Research. Journal of Anhui Agri. Sci. 36(22), 9779–9782 (2008)
3. Hoerl, A.E., Kenard, R.W., Baldwin, K.F.: Ridge regression: Some simulation. Communications in Statistics 4(2), 105–123 (1975)
4. Liu, H.-C., Tu, Y.-C., Lin, W.-C., Chen, C.C.: Choquet integral regression model based on L-Measure and γ-Support. In: Proceedings of 2008 International Conference on Wavelet Analysis and Pattern Recognition (2008)
5. Liu, H.-C.: Extensional L-Measure Based on any Given Fuzzy Measure and its Application. In: Proceedings of 2009 CACS International Automatic Control Conference, National Taipei University of Technology, Taipei Taiwan, November 27-29, pp. 224–229 (2009)
6. Liu, H.-C., Ou, S.-L., Cheng, Y.-T., Ou, Y.-C., Yu, Y.-K.: A Novel Composition Forecasting Model Based on Choquet Integral with Respect to Extensional L-Measure. In: Proceedings of The 19th National Conference on Fuzzy Theory and Its Applications (2011)
7. Choquet, G.: Theory of capacities. Annales de l'Institut Fourier 5, 131–295 (1953)
8. Wang, Z., Klir, G.J.: Fuzzy Measure Theory. Plenum Press, New York (1992)
9. Sugeno, M.: Theory of fuzzy integrals and its applications. unpublished doctoral dissertation, Tokyo Institute of Technology, Tokyo, Japan (1974)
10. Zadeh, L.A.: Fuzzy Sets as a Basis for Theory of Possibility. Fuzzy Sets and Systems 1, 3–28 (1978)

Aggregating Multiple Robots with Serialization

Shota Sugiyama[1], Hidemi Yamachi[2],
Munehiro Takimoto[3], and Yasushi Kambayashi[2]

[1] Graduate School of Computer and Information Engineering
[2] Department of Computer and Information Engineering,
Nippon Institute of Technology 4-1 Gakuendai, Miyashiro-machi, Minamisaitama-gun,
Saitama, 345-8501 Japan
c1065302@cstu.nit.ac.jp
{yamachi,yasushi}@nit.ac.jp
[3] Department of Information Sciences, Tokyo University of Science
2641 Yamazaki, Noda 278-8510 Japan
mune@cs.is.noda.tus.ac.jp

Abstract. This paper presents the design of an intelligent cart system to be used in a typical airport. The intelligent cart system consists of a set of mobile software agents to control the cart and provides a novel method for alignment. If the carts gather and align themselves automatically after being used, it is beneficial for human workers who have to collect them manually. To avoid excessive energy consumption through the collection of the carts, in the previous study, we have used ant colony optimization (ACO) and a clustering method based on the algorithm. In the current study, we have extended the ACO algorithm to use the vector values of the scattered carts in the field instead of mere location. We constructed a simulator that performs ant colony clustering using vector similarity. Waiting time and route to the destination of each cart are made based on the cluster created this way. These routes and waiting times are conveyed by the agent to each cart, while making them in rough lines. Because the carts are clustered by the similarity of vectors, we have observed that several groups have appeared to be aligned. The effectiveness of the system is demonstrated by constructing a simulator and evaluating the results.

Keywords: Mobile software agent, Autonomous system, Ant colony optimization, Ant colony clustering, Simulation.

1 Introduction

When we pass through terminals of an airport, we often see carts scattered in the walkway and laborers manually collecting them one by one. It is a laborious task and not a fascinating job. It would be much easier if carts were roughly gathered in any way before the laborers begin to collect them. Multi-robot systems have made rapid progress in various fields, and the core technologies of multi-robots systems are now easily available [1]. Therefore, it is possible to make each cart have minimum

J.-S. Pan, S.-M. Chen, N.T. Nguyen (Eds.): ACIIDS 2012, Part I, LNAI 7196, pp. 177–186, 2012.

intelligence, making each cart an autonomous robot. We realize that for such a system cost is a significant issue and we address one of those costs, the power source. A big powerful battery is heavy and expensive; therefore such intelligent cart systems with small batteries are desirable. Thus energy saving is an important issue in such a system [2].

We use mobile agents to drive the carts that are placed in various locations to the quasi-optimal destinations. So, through Ant Colony Optimization (ACO) with a simulation, we choose the method of determining the destination of each cart. Because clustering with the ACO's position can be determined from the optimal set using ant agents that exist only on the simulator [3]. Earlier studies have succeeded in making some beautiful groups [4] [5]. However, because we collected carts this way it looks overindulgent, and a good alignment is hard. So we decided to do a better alignment of the carts at the same time of their collection. In the improved method, Ant Colony Clustering (ACC) has been used to calculate the best pheromone set position, by making a degree of similarity in the vector; we should be able to align the orientation of the cart in a cluster.

The structure of the balance of this paper is as follows. In the second section, we describe the background. In the third section, we describe the agent system that performs the arrangement of the intelligent carts. The agent system consists of several static and mobile agents. The static agents interact with the users and compute the ACC algorithm, and other mobile agents gather the initial positions of the carts and distribute the assembly positions. The fourth section describes the simulation field, and the initial coordinates and directions of the scattered carts in the field. The fifth section describes the ant colony clustering (ACO) that uses vector value of each cart has. The algorithm draws the carts that have similar vector values so that carts that are roughly facing the same direction get together. The sixth section describes the simulator and demonstrates the effectiveness of our algorithm through numerical experiments. Finally, in the seventh section we summarize the work and discuss future research directions.

2 Background

Kambayashi and Takimoto have proposed a framework for controlling intelligent multiple robots using higher-order mobile agents [1][2]. The framework helps users to construct intelligent robot control software by migration of mobile agents. Since the migrating agents are higher-order, the control software can be hierarchically assembled while they are running. Dynamically extending control software by the migration of mobile agents enables them to make base control software relatively simple, and to add functionalities one by one as they know the working environment. Thus they do not have to make the intelligent robot smart from the beginning or make the robot learn by itself. They can send intelligence later as new agents. Even though they demonstrate the usefulness of the dynamic extension of the robot control software by using the higher-order mobile agents, such higher-order property is not necessary in our setting. They have implemented a team of cooperative search robots to show the effectiveness of their framework, and demonstrated that their framework contributes to energy saving of multiple robots [2]. They have achieved significant saving of energy.

Deneubourg has formulated the biology inspired behavioral algorithm that simulates the ant corps gathering and brood sorting behaviors [6]. Lumer has improved Deneubourg's model and proposed a new simulation model that is called Ant Colony Clustering [7]. His method could cluster similar objects into a few groups. Kambayashi et al have improved Lumer's model and proposed a multi-robot control by using ACC [4]. This method does not cluster similar objects but cluster nearby objects by using pheromone. Kambayashi et al have designed an intelligent cart system based on the proposed method [8]. As mentioned above, previous studies could collect mobile robots roughly using ACC. In this paper, we try to align the mobile robots while collecting them.

3 System Model

Our system model consists of carts and a few kinds of static and mobile software agents. All the controls for the mobile carts as well as ACC computation performed in the host computer are achieved through the static and mobile agents. They are: 1) user interface agent (UIA), 2) operation agents (OA), 3) position collecting agent (PCA), 4) clustering simulation agent (CSA), and 5) driving agents (DA). All the software agents except UIA and CSA are mobile agents. A mobile agent traverses carts scattered in the field one by one to collect their coordinates. After receiving the assembly positions computed by a static agent, many mobile agents migrate to the carts and drive them to the assembly positions. Fig. 1 shows the interactions of the cooperative agents to control an intelligent cart. The followings are details of each agent:

1) User Interface Agent (UIA): The user interface agent (UIA) is a static agent that resides on the host computer and interacts with the user. It is expected to coordinate the entire agent system. When the user creates this agent with a list of IP addresses of the intelligent carts, UIA creates PCA and passes the list to it.
2) Operation Agent (OA): Each cart has at least one operation agent (OA). It has the task that the cart on which it resides is supposed to perform. Each intelligent cart has its own OA. Currently all operation agents (OA) have a function for collision avoidance and a function to sense RFID tags embedded in the floor carpet to detect its precise coordinates in the field.
3) Position Collecting Agent (PCA): A distinct agent called position collecting agent (PCA) traverses carts scattered in the field one by one and collects their coordinates. PCA is created and dispatched by UIA. Upon returning to the host computer, it hands the collected coordinates to the clustering simulation agent (CSA) for ACC.

Clustering Simulation Agent (CSA): The host computer houses the static clustering simulation agent (CSA). This agent actually performs the ACC algorithm by using the coordinates collected by PCA as the initial positions, and produces the quasi-optimal assembly positions of the carts, and then performs yet

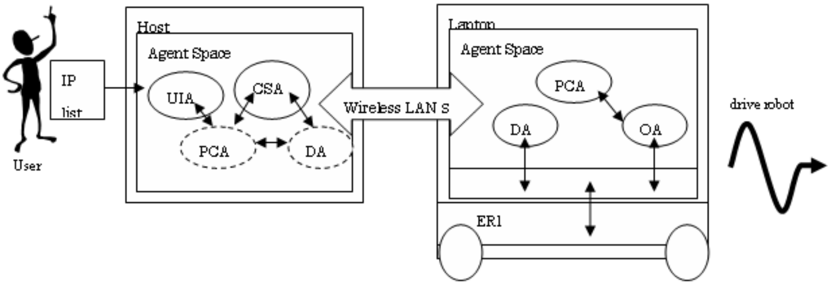


Fig. 1. Cooperative agents to control an intelligent cart

another simulation to produce instructions each cart follows to reach its assigned goal position. Upon terminating the simulation and producing the procedures for all the intelligent carts, CSA creates a number of driving agents (DA).

5) Driving Agent (DA): The quasi-optimal arrangement coordinates, as well as procedures to reach them, produced by the CSA are delivered by driving agents (DA). One driving agent is created for each intelligent cart, and it contains the set of procedures for the cart. The DA drives its intelligent cart to the designated assembly position. DA is the intelligent part of the intelligent cart.

OA detects the current coordinate of the cart on which it resides. Each cart has its own IP address and UIA hands in the list of the IP addresses to PCA. First, PCA migrates to an arbitrary cart and starts hopping between them one by one. It communicates locally with OA, and writes the coordinates of the cart into its own local data area. When PCA gets all the coordinates of the carts, it returns to host computer. Upon returning to the host computer, PCA creates CSA and hands in the coordinate data to CSA which computes the ACC algorithm.

The current implementation employs RFID (Radio Frequency Identification) to get precise coordinates [9]. We set RFID tags in a regular grid shape under the floor carpet tiles. The tags we chose have a small range so that the position-collecting agent can obtain fairly precise coordinates from the tags. Also, the cart has a basic collision avoidance mechanism using infrared sensors.

CSA is the other static agent and its sole role is ACC computation. When CSA receives the coordinate data of all the carts, it translates the coordinates into coordinates for simulation, and then performs the clustering. When CSA finishes the computation and produces a set of assembly positions, it then creates the set of procedures for autonomous cart movements.

Then CSA creates DA that conveys the set of procedures to the intelligent carts as many. Each DA receives its destination IP address from PCA, and the set of procedures for the destination cart, and then migrates to the destination cart. Each DA has a set of driving procedures that drives its assigned cart to the destination, while it avoids collision. OA has the basic collision detection and avoidance procedures, and DA has task-specific collision avoidance procedures.

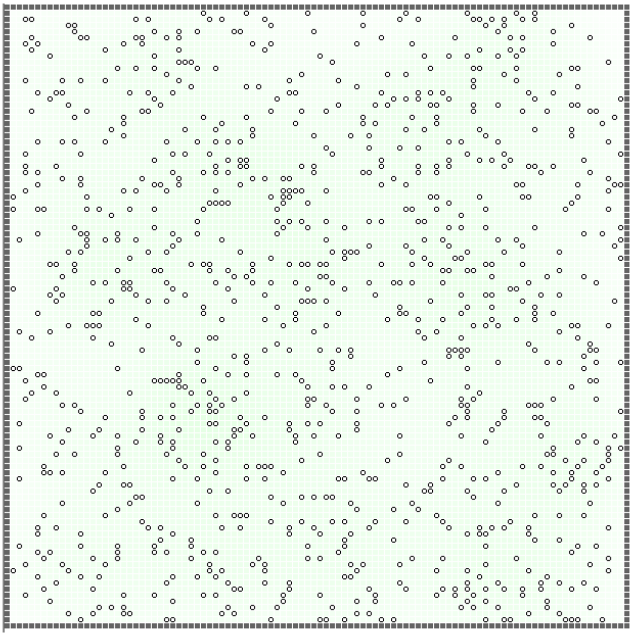


Fig. 2. Simulation Field

4 Initial Coordinates and Directions

The purpose of this study is to assemble robot carts scattered in a field into several short lines. In order to achieve this task, we have developed an extended ACO algorithm to collect robot carts that have similar vector values, i.e. facing the similar directions. In order to develop and validate our algorithm, we have developed a simulator. In this and the following sections concentrate the extended ACO algorithm we have developed through the simulator. Since the idea is general, we henceforth use “object” instead of “cart.”

Simulation field is the n × n grid of two-dimensional grid (Fig. 2). Since the edge of the field is surrounded by walls so that ants and artificial objects do not go out of the field. The initial position vectors of the objects in the simulator are determined randomly each time the simulator starts. The number of objects and field size of the simulator can be configured by the user.

5 Extended Ant Colony Clustering

ACO is a swarm intelligence-based method and a multi-agent system that exploits artificial stigmergy for the solution of combinatorial optimization problems. Ant colony clustering (ACC) is an ACO specialized for clustering objects. The idea is inspired by the collective behaviors of ants, and Deneubourg formulated an algorithm that simulates the ant corps gathering and brood sorting behaviors [6].

In traditional systems ACO, artificial ants that are able to follow the same path of other artificial ants leave pheromone signal. Our extended ACO system uses pheromone as the similarity of vector values. Pheromone is generated from objects, diffused into the surroundings of the objects. The farther away from the object, the weaker the property becomes (Fig. 3).

Artificial ants are moving in random directions every eight discrete time, and they pick up the weak vector objects. In addition, an artificial ant has an object, if the vector of a nearby object that has high similarity with the object, the artificial ant has a habit of putting objects there. When an artificial ant had placed an object, the object vector is synthesized with vectors of surrounding objects into one vector.

Artificial ant behavior in the ACC of this study is determined by the equation (1), (2), (3).

Equation (1), an artificial ant has discovers an object *i*, the ant determines whether the expression suggests to pick up the object or not. Kp is the size of the norm that forms the cluster. *f (i)* is the magnitude of the norm of object *i* found. The norm is the magnitude of the vector. When an artificial ant has discovered the value of the norm of the object, it compares the Kp. If Kp is less than the norm of the object the artificial ant found, it pick up the object.

Equation 2, when the newly discovered object *j* an artificial ant has an object *i*, is an expression determines whether the ant places the object *i*. *Vs* is a constant for the determination of similarity. *cos (i, j)* represents the similarity of objects *i* and *j*. If an artificial ant finds a new object and that has an object, artificial ants compares the similarity of vector object and found object. If the value *Vs* exceeds the predetermined degree of similarity obtained at this time, the ant put the object.

Equation (3) is used when a determination is made in Equation (2) is used to compute the similarity between an object and a vector *i* of *j*. This value is represented by 0 and 1. We let vanishingly low affinity be 0 and we let high affinity be 1. <*x, y*> expresses the dot product of y and *x*, and | | *x* | | is the norm of x. In determining whether to place the object in Equation 2 is obtained by this similarity.

When creating a group of several objects with similar vectors, our goal is to move each cart the shortest distance to generate a cluster. In order to achieve this goal, the artificial ants have the following regulatory action.

1. An artificial ant finds a cluster of objects and the cluster has more than a certain norm, the artificial ant avoids picking up an object from the cluster. This number can be updated.
2. An artificial ant with an object cannot find a cluster with the strength of the norm, it moves randomly. Direction of movement of artificial ants that time is affected by the vector of the object it has.
3. An artificial ant with an object cannot find any cluster within a certain walking distance; it put the object back into the current position and starts the search again.

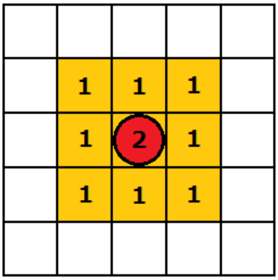


Fig. 3. Distribution of pheromone

$$P_{pick}(i) = \begin{cases} 1, & if \quad f(i) < Kp \\ 0, & otherwise \end{cases} \tag{1}$$

$$P_{drop}(i) = \begin{cases} 1, & if \quad Vs < \cos(i, j) \\ 0, & otherwise \end{cases} \tag{2}$$

$$\cos\theta = \frac{\langle x, y \rangle}{\|x\| . \|y\|} \tag{3}$$

Based on the above rules, artificial ants are aligned to form a cluster of objects with similar vector values the same as the phase 1 so that the ants need not to carry them for a long distance. In order to implement the phase 1, the system locks the objects with the norm above a certain feature so that any ants do not pick them up. Once clustering is emerging, to update the size of the norm to be fixed so that artificial ants can bring what were previously fixed. This is the rule to form a cluster with a larger norm.

Features of 2, if an artificial ant cannot sense pheromone around it, it attempts to discover pheromone by moving at random to one of the 8 squares around it. Consideration of the costs incurred when changing the orientation of the cart, and then to move forward toward the object as possible vectors (Fig. 4).

Feature of the 3 is the ability to reset the movement of artificial ants become too expensive to move. The travel distance of each object is limited. This is intended to eliminate the energy loss caused by unnecessary movement. Once the object is moved to the left to right to find a cluster with a similar vector, it is loss energy. In order to prevent such a situation, this feature is essential.

Several large clusters with the same vector values should emerge by repeating the above rules eventually.

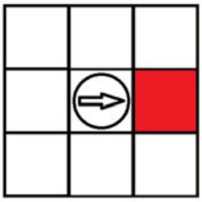


Fig. 4. Examples of artificial ants advancing towards the object vector

6 Flow of the Simulation and Experimental Results

In this section, we report the results of the experiments we have conducted through the simulator. Simulator is started with randomly placed objects in the field. The number of artificial ants is determined by the user. Each artificial ant checks pheromones around it at the every discrete time step to receive an order to commence. Artificial ants, based on the value of this pheromone, find the object, pick it up, take action and place it. Clusters of objects are formed based on the similarity of the vector values. The newly formed cluster synthesizes pheromones of the object to form a larger cluster. These clusters will update the value of the norm so that no artificial ants picked up the objects that form a cluster. The simulator repeats this procedure until the target number of clusters is formed.

When a certain number of clusters are formed, the simulator will start another simulation. This simulation is for calculating the set position and movement path as well as waiting timing of each object. Each object will move one square at each step by the shortest route to their meeting place. If the object tries to move the position where another object is, it waits to avoid a collision. Each object on the display leaves a trajectory when it moves in the simulator so that the user of simulator can see how the objects have moved at a glance.

The simulator evaluates the simulation results by using the sum of the distances of objects and the number of clusters and the similarity of the vectors in the cluster. The less the number of clusters and the sum of the distances, the higher it rates. If the similarity of the vector is closest to 1 rated high. Objects belonging to each cluster must lower the sum of the distances and also cluster similarity is high. The vector value formed by clustering with the similarity affects behavioral characteristics of ants with objects.

In this experiment, the field size was 100 × 100, is there was 100 objects and 100 artificial ants. Experiment was set up this way (Fig. 5).

Approximately 70% of the objects within each cluster are in the same direction. The resulted clusters generated by the similarity of vector values show the experiments are succeeded. Also, the cost of forming clusters of objects with similar vector values is not different from that of without vector values. We can say that the moving cost is successfully suppressed (Table 1). The experiments suggest that our method is useful.

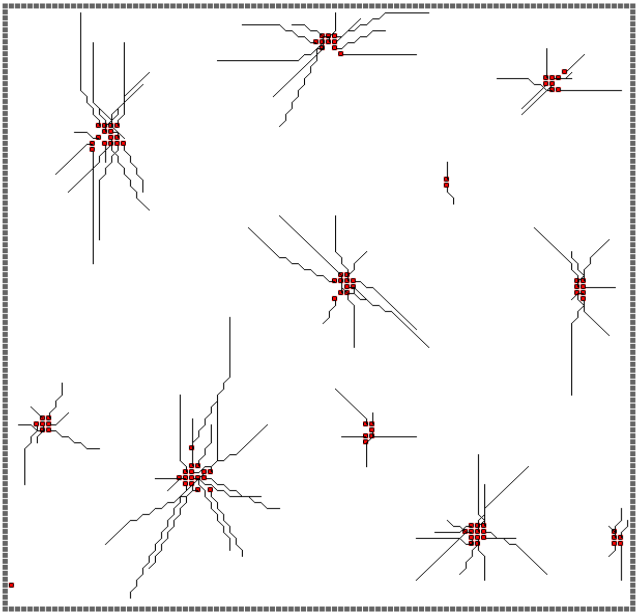


Fig. 5. Simulation Results

Table 1. Comparison of this study and previous research

Old				New			
No	C_No	Cost	AveCost	No	C_No	Cost	AveCost
1	10	934	8.50	1	9	1029	9.22
2	11	929	8.18	2	8	1077	9.88
3	10	905	8.60	3	12	881	8.17
4	8	1046	9.63	4	7	1099	11.00
5	11	907	7.73	5	11	916	7.73
Ave	10	944.2	8.53	Ave	9.4	1000.4	9.20

7 Summary

This paper present an algorithm to make clusters of objects based on the similarity of the vector values of objects. The algorithm is an extension of the ant colony optimization. This algorithm is to be used in a framework for controlling the robot carts used in an airport. Since the similarity is the vector value, the robots in the formed clusters must tend to have similar directions, i.e. facing the same direction.

This feature must greatly reduce the manual labor work, when it is implemented in real environment.

We have constructed a simulator to demonstrate the effectiveness of our algorithm. Approximately 70% of the objects within each cluster are facing in the same direction. The resulted clusters that are generated by the vector is said to have succeeded. We are re-implanting the algorithm to adjust the detection range of the pheromone and pheromone concentration ratio so that more precise alignment can be achieved.

As the next step, we plan not only to align the direction of the robots but also to serialize the formed robots so that we can provide more benefits to the cart collection.

References

1. Kambayashi, Y., Takimoto, M.: Higher-Order Mobile Agents for Controlling Intelligent Robots. International Journal of Intelligent Information Technologies 1(2), 28–42 (2005)
2. Takimoto, M., Mizuno, M., Kurio, M., Kambayashi, Y.: Saving Energy Consumption of Multi-robots Using Higher-Order Mobile Agents. In: Nguyen, N.T., Grzech, A., Howlett, R.J., Jain, L.C. (eds.) KES-AMSTA 2007. LNCS (LNAI), vol. 4496, pp. 549–558. Springer, Heidelberg (2007)
3. Kambayashi, Y., Yamachi, H., Takimoto, M.: A Search for Practical Implementation of the Intelligent Cart System. In: Congress on Computer Applications and Computational Science, pp. 895–898 (2010)
4. Kambayashi, Y., Ugajin, M., Sato, O., Tsujimura, Y., Yamachi, H., Takimoto, M., Yamamoto, H.: Integrating Ant Colony Clustering Method to a Multi-Robot System Using Mobile Agents. Industrial Engineering and Management System 8(3), 181–193 (2009)
5. Kambayashi, Y., Yamachi, H., Tsujimura, Y.: A Search for Efficient Ant Colony Clustering. In: Asia Pacific Industrial Engineering and Management Systems Conference, pp. 591–602 (2009)
6. Deneubourg, J., Goss, S., Franks, N., Sendova-Franks, A., Detrain, C., Chretien, L.: The Dynamics of Collective Sorting: Robot-Like Ant and Ant-Like Robot. In: First Conference on Simulation of Adaptive Behavior: From Animals to Animats, pp. 356–363. MIT Press (1991)
7. Lumer, E.D., Faiesta, B.: Diversity and adaptation in populations of clustering ants, from animals to animats 3. In: 3rd International Conference on the Simulation of Adaptive Behavior, pp. 501–508. MIT Press, Cambridge (1994)
8. Kambayashi, Y., Harada, Y., Sato, O., Takimoto, M.: Design of an Intelligent Cart System for Common Airports. In: 13th IEEE International Symposium Consumer Electronics, CD-ROM (2009)

A Multi-attribute and Multi-valued Model for Fuzzy Ontology Integration on Instance Level*

Hai Bang Truong[1] and Ngoc Thanh Nguyen[1,2]

[1] University of Information Technology, VNU-HCM, Vietnam
bangth@uit.edu.vn
[2] Wroclaw University of Technology, Poland
ngoc-thanh.nguyen@pwr.edu.pl

Abstract. Fuzzy ontology are often more useful than non-fuzzy ontologies in knowledge modeling owing to the possibility for representing the incompleteness and uncertainty. In this paper we present an approach to fuzzification on the instance level of ontology using multi-value and multi-attribute structure. A consensus-based method for fuzzy ontology integration is proposed.

1 Introduction

In an integration process most often one of the following aspects is realized:

- Several objects are merged to give a new object best representing them (merging aspect)
- Several objects create a "union" acting as a whole (alignment aspect)
- Several objects are corresponded with each other (mapping aspect).

These aspects are most important and most popular in integration processes of information systems.

In the knowledge management field, an integration task most often refers to a set of elements of knowledge with the same kind of semantic structures, the aim of which is based on determining an element best representing the given. The kinds of structures mean, for example relational, hierarchical, logical etc. The words "best representing" mentioned above refer to the merging aspect and mean the following criteria for integration:

- All data included in the objects to be integrated should be included in the result of integration. Owing to this criterion all pieces of information included in the component elements will appear in the integration result.
- All inconsistencies of elements to be integrated should be resolved. It often happens that referring to the same subject different elements contain inconsistent data. Such situation is called an inconsistency. The integration result should not contain inconsistency, that is integrity constraints should be fulfilled.

* This work was partially supported by Vietnam National Foundation for Science and Technology Development (NAFOSTED) under grant number 102.01-2011.10 (2011-2013) and by Polish Ministry of Science and Higher Education under grant no. N N519 407437 (2009-2012).

J.-S. Pan, S.-M. Chen, N.T. Nguyen (Eds.): ACIIDS 2012, Part I, LNAI 7196, pp. 187–197, 2012.

Integration tasks are very often realized for database integration or knowledge integration processes. Ontology integration is a special case of the second case. Ontologies have well-defined structure and it is assumed that the result of ontology integration is also an ontology. Therefore, the first and second criteria are most popular.

It seems that satisfying the first criterion is simple since one can creating a new ontology making the sum of all sets of concepts, relations and axioms from component ontologies. However, it is not always possible because of the following reasons:

- Occurrence of all elements in the integration result may contain inconsistency in the sense that some of the component ontologies may be in conflict and this conflict will be moved to the integration result.
- Including all elements in the integration result may cause conflicts of the relations between ontology concepts.

Satisfying the second criterion is based on solving conflicts, for example, by using consensus methods [15].

Similarly like for non-fuzzy ontologies [14], conflicts between fuzzy ontologies may also be considered on the following levels:

- Conflicts on concept level: The same concept has different structures in different ontologies.
- Conflicts on relation level: The relations between the same concepts are different in different ontologies.
- Conflicts on instance level: The same instance has different descriptions in different concepts or ontologies.

The subject of this paper is working out algorithms for ontology integration on instance level. In the next section we present the structure of ontology on instance level and the definition of integration. In Section 3 a model for fuzzy instance integration using multi-value and multi-attribute approach is included. Section 4 includes a set of postulates for integration. In Section 5 an algorithm for fuzzy ontology integration on instance level are presented and finally in Section 6 a brief overview of related works is included.

2 Multi-value and Multi-attribute Structure for Fuzzy Ontology Representation

2.1 Definition of Fuzzy Ontology

The basis for determining an ontology is a real world ($\boldsymbol{A}$, $\boldsymbol{V}$) where $\boldsymbol{A}$ is a finite set of attributes describing the domain and $\boldsymbol{V}$ – the domain of $\boldsymbol{A}$, that is $\boldsymbol{V}$ is a set of attribute values, and $\boldsymbol{V} = \bigcup_{a\in \boldsymbol{A}} V_a$ (where V_a is the domain of attribute a). In the previous work [16] we have presented the following definition of fuzzy ($\boldsymbol{A}$,$\boldsymbol{V}$)-based ontology:

$$\textit{Fuzzy ontology} = (\boldsymbol{C}, \boldsymbol{R}, \boldsymbol{Z})$$

where:

- $\boldsymbol{C}$ is the finite set of concepts. A concept of a fuzzy ontology is defined as a triple:

$$\textit{concept} = (c, A^c, V^c, f^c)$$

where c is the unique name of the concept, $A^c \subseteq \boldsymbol{A}$ is a set of attributes describing the concept and $V^c \subseteq \boldsymbol{V}$ is the attributes' domain: $V^c = \bigcup_{a \in A^c} V_a$ and f^c is a fuzzy function:

$$f^c: A^c \to [0,1]$$

representing the degrees to which concept c is described by attributes from set A^c. Triple (A^c, V^c, f^c) is called the *fuzzy structure* of concept c.

- $\boldsymbol{R}$ is a set of fuzzy relations between concepts, $\boldsymbol{R} = \{R_1, R_2,\ldots, R_m\}$ where

$$R_i \subseteq \boldsymbol{C} \times \boldsymbol{C} \times (0, 1]$$

for $i = 1, 2,\ldots,m$. A relation is then a set of pairs of concepts with a weight representing the degree to which the relationship should be. We assume that within a relation R_i in an ontology a relationship can appear between two concepts only with one value of weight, that is if $(c, c', v) \in R_i$ and $(c, c', v') \in R_i$ then $v = v'$.

- $\boldsymbol{Z}$ is a set of constraints representing conditions on concepts and their relationships. In this paper we do not deal with them.

Note that in the above definitions there is no reference to fuzzy aspect of instances. It turned out that even in a non-fuzzy ontology where the knowledge about concepts and their relations are complete and certain, there may appear some problem with classifying instances to concepts because, for example, some attribute values for an instances may be incomplete or unknown. For solving this problem, it seems that the multi-value structure is very useful.

Here we will present a concept for representing the uncertainty and incompleteness of instance description by means of a multi-value and multi-attribute structure.

Each attribute $a \in \boldsymbol{A}$ has a domain as a set V_a of elementary values. A value of attribute a may be a subset of V_a as well as some element of V_a. Set 2^{V_a} is called the *super domain* of attribute a. For $B \subseteq \boldsymbol{A}$ let's denote

$$V_B = \bigcup_{b \in B} V_b \text{ and } \overline{2}^B = \bigcup_{b \in B} 2^{V_b}.$$

Definition 1. *A fuzzy instance of a concept c is described by the attributes from set A^c with values from set 2^{V_X} (for $X = A^c$) and is defined as a pair:*

$$instance = (i, v)$$

where i is the unique identifier of the instance in world $(\boldsymbol{A}, \boldsymbol{V})$ and v is the value of the instance and is a tuple of type A^c and can be presented as a function:

$$v: A^c \to \overline{2}^{A^c}$$

such that $v(a) \in 2^{V_a}$ for all $a \in A^c$.

Value v is also called a description of the instance within a concept. We can note that an attribute value in an instance is not a single value, but a set of values. This is because it is not certain which value is proper for the attribute. Owing to this, the fuzziness can be represented. Note that in this case the fuzziness of attribute values is not represented by a number, but by a set of values. The interpretation is that the proper value of an attribute in an instance is an element of the set, it is not known which one.

For fuzzy instances the *Instance Integration Condition* (IIC) should be satisfied. That is the descriptions of the same instance in different concepts should be consistent. However, the same instance may belong to different concepts and may have different descriptions. The following condition should be satisfied:

Let instance i belong simultaneously to concept c with description (i, v) *and to concept* c' *with description* (i, v'). *If* $A^c \cap A^{c'} \neq \emptyset$ *then there should be* $v(a) \cap v'(a) \neq \emptyset$ *for each* $a \in A^c \cap A^{c'}$.

2.2 A Multi-value and Multi-attribute Structure for Fuzzy Instance Representation

For a real world $(\boldsymbol{A}, \boldsymbol{V})$ we define the following notions [12]. Let $B \subseteq \boldsymbol{A}$.

- A *complex tuple* (or *tuple* for short) of type B is a function

  $$r: B \to \overline{2}^{V_B}$$

  such that $r(a) \subseteq V_a$ for all $a \in B$. Instead of $r(a)$ we will write r_a and a tuple of type B will be written as r_B.

  A tuple r of type B may be written as a set:

  $$r = \{(a, r_a): a \in B\}.$$

- An *elementary tuple* of type B is a function

  $$r: B \to V_B$$

  such that $r(a) \in V_a$ for all $a \in B$. If $V_a = \emptyset$ then $r(a) = \varepsilon$, where symbol ε represents a special value used for case when the domain is empty.

  The set of all elementary tuples of type B is denoted by $E\text{-}TU(B)$.

  An elementary tuple r of type B may also be written as a set:

  $$r = \{(a, r_a): a \in B\}.$$

- By symbol ϕ we denote the set of all empty tuples, i.e. whose all values are empty. By symbol ϕ_{E} we denote the set of all empty elementary tuples, i.e. whose all values are equal ε.
- By symbol θ we denote the set of all partly empty tuples, i.e. in which at least one value is empty. Expression $r \in \theta$ will mean that in tuple r at least one attribute value is empty and expression $r \notin \theta$ will mean that in tuple r all attribute values are not empty. Of course we have $\phi \subset \theta$. By symbol θ_E we denote the set of all partly empty elementary tuples.
- The sum of two tuples r and r' of type B is a tuple r'' of type B such that $r''_a = r_a \cup r'_a$ for each $a \in B$. This operation is written as

  $$r'' = r \cup r'.$$

  More generally, a sum of two tuples r and r' of types B and B', respectively, is a tuple r'' of type $B'' = B \cup B'$ such that

$$r'' = \begin{cases} r_b \cup r_{b'} & \text{for } b \in B \cap B' \\ r_b & \text{for } b \in B \setminus B' \\ r_{b'} & \text{for } b \in B' \setminus B \end{cases}$$

- The product of two tuples r and r' of type B is also a tuple r'' of type B such that $r''_a = r_a \cap r'_a$ for each $t \in B$. This operation is written as

  $$r'' = r \cap r'.$$

  More generally, a product of two tuples r and r' of types B and B', respectively, is a tuple r'' of type $B'' = B \cap B'$ such that $r''_a = r_a \cap r'_a$ for each $a \in B \cap B'$. This operation is written as:

  $$r'' = r \cap r'.$$

- Let $r \in TU(B)$ and $r' \in TU(B)$ where $B \subseteq B'$, we say that tuple r is included in tuple r', if and only if $r_a \subseteq r'_a$ for each $a \in B$. This relation is written as

  $$r \prec r'.$$

Now we deal with the way for measuring the distance between attribute values in complex tuples. A distance is measured between two values of the same attribute.

Let $a \in \boldsymbol{A}$ be an attribute, as the distance function for values of attribute a we understand a function:

$$d_a : 2^{V_a} \times 2^{V_a} \to [0, 1].$$

We assume that V_a is a finite set, and let $card(V_a) = N$. In [12] there have been defined two kinds of distance functions. They are the following:

A. Functions Minimizing Transformation Costs

The general idea of this kind of distance functions is relied on determining the distance between two sets as the minimal cost needed for transforming one set into the other. A set X is transformed into a set Y by means of such operations as *Adding*, *Removing* and *Transformation*, which are used for elements of set X to obtain set. The following cost functions have been defined:

- Function E_{AR} representing the cost for adding (or removing) of an elementary value to (or from) a set;
- Function E_T representing the cost for transformation of one elementary value into another.

Definition 2. *By distance function minimizing the transformation cost we understand the following function:*

$$\delta_a : 2^{V_a} \times 2^{V_a} \to [0,1]$$

which for a pair of sets X, $Y \subseteq V_a$ assigns a number

$$\delta_a(X,Y) = \frac{T(X,Y)}{\sum_{x\in V_a} E_{AR}(x)}$$

where T(*X*,*Y*) *represents the minimal cost needed for transforming set X into set Y.*

B. Functions Reflecting Element Contribution in The Distance

This kind of functions is based on determining the value of contribution of each element of set V_a in the distance between two subsets of this set. Of course, the contribution of an element depends on the sets between which the distance is measured.

Let $a \in \boldsymbol{A}$, we define the following 3–argument contribution function:

$$S_a\colon 2^{V_a} \times 2^{V_a} \times V_a \to [0, 1]$$

such that value $S_a(X,Y,z)$ represents the contribution of element z in the distance between two sets X and Y.

The distance between two sets should now be defined as the sum of contributions of elementary values referring to these sets, by means of the following function:

$$\rho_a\colon 2^{V_a} \times 2^{V_a} \to [0,1].$$

Definition 3. *For any sets X*, $Y \subseteq V_a$ *their distance* $\rho_a(X,Y)$ *is equal to*

$$\rho_a(X,Y) = \frac{N}{2N-1}\sum_{z\in V_a} S_a(X,Y,z)\,.$$

Owing to the above defined functions one can define a function d between tuples as a combination of them.

3 Problem of Fuzzy Instance Integration

We mention that the inconsistency on instance level is based on the fact that the same instance is described differently in different concepts from different ontologies. This is because of the fuzzy feature of instance descriptions. We have assumed that within an ontology the *Instance Integration Condition* must be satisfied.

For integration of ontologies on instance level consensus methods seem to be very useful. Different criteria, structures of data and algorithms for consensus choice have been worked out [9], [10]. Here we assume that we have to deal with a set of versions of the same instance. A version of an instance (i, x) can be represented by the value x. The task is to determine a version which best represents the given versions.

For this kind of consistency the consensus problem can be defined as follows:

Given a set of instance descriptions

$$X = \{(i, v_1),\ldots, (i, v_n)\}$$

where v_i *is a tuple of type* $A_i \subseteq \boldsymbol{A}$, *that is* $v_i\colon A_i \to V_i$ *for* $i = 1,\ldots, n$ *and* $V_i = \bigcup_{a\in A_i} V_a$ *one should find a description* (i, v) *of some type*, *which best represents the given values.*

As shown in [12], it has turned out that in this case the best criterion for determining value v is the following:

$$\sum_{i=1}^{n} d(v, v_i) = \min_{v' \in TYPE(A)} \sum_{i=1}^{n} d(v', v_i)$$

where $A = \bigcup_{i=1}^{n} A_i$ and d is a distance function between tuples.

Algorithms for determining consensus for the above problem can be found in [12]. These algorithms are dependent on the structure of attribute values. The value v which is determined by one of these algorithms can be assumed to be the value of instance i in the final ontology.

In many cases integration task is not based on consensus determining. As shown above, consensus calculation is most often based on resolution of an optimization problem. However, note that optimization is only one of possible conditions for integration result. As presented in the next section, we define several criteria for instance description integration. These criteria represent some intuitive and rational requirements for determining a description proper for an instance which have different descriptions in different in ontologies, or in the same ontology, but in different concepts. We define the problem of instance descriptions integration as follows:

Given a set of instance descriptions

$$X = \left\{ t_i \in TU(T_i)\text{: } T_i \subseteq \boldsymbol{A} \textit{ for } i = 1, 2, \ldots, n \right\}$$

one should determine a tuple t^* of type $T^* \subseteq \boldsymbol{A}$ which best represents the given tuples.

Tuple t^* is called an integration of instance descriptions. Notice that the integration problem is different from a consensus problem in that there is no assumption that the tuples representing the agent knowledge states are of the same type, as is in the consensus problem [13].

4 Postulates for Fuzzy Instance Integration

In order to integrate different versions of the same instance, we propose some criteria. These criteria are used in order to verify the fulfillment of minimal requirements for the integration process. Similarly as in work [12] here we define the following criteria for fuzzy instance integration:

C_1. *Instance Closure*:

Tuple t^* should be included in the sum of given tuples, that is

$$t^* \prec \bigcup_{i=1}^{n} t_i$$

C_2. *Instance Consistency*

The common part of the given tuples should be included in tuple r^*, that is

$$\bigcap_{i=1}^{n} t_i \prec t^*$$

C_3. *Instance Superiority*

If sets of attributes T_i $(i = 1,2,\ldots,n)$ *are disjoint with each other then*

$$t^* = \left[\bigcup_{i=1}^{n} t_i\right]_{T^*}$$

where $\left[\bigcup_{i=1}^{n} t_i\right]_{T^*}$ *is the sum* $\bigcup_{i=1}^{n} t_i$ *restricted to attributes from set T*.*

C_4. *Maximal similarity*

Let d_a *be a distance function values of attribute* $a \in \boldsymbol{A}$ *then the difference between integration t* and the profile elements should be minimal in the sense that for each* $a \in T^*$ *the sum*

$$\sum_{r \in Z_a} d(t^*_a, r)$$

where

$$Z_a = \{r_{ia}\text{: } r_{ia} \textit{ is definite}, i = 1, 2,\ldots, n\}$$

should be minimal.

5 Algorithms for Fuzzy Instance Integration

Below we present an algorithm for instance integration which satisfies criteria C_1, C_2, C_3 and C_4. The idea of this algorithm is based on determining consensus for each of attributes and next completing their values for creating the integration.

Algorithm

Input: Set of *n* descriptions of an instance:

$$X = \{r_i \in TU(T_i)\text{: } T_i \subseteq \boldsymbol{A} \text{ for } i = 1, 2,\ldots, n\}$$

and distance functions d_a for attributes $a \in \boldsymbol{A}$.

Output: Tuple t^* of type $T^* \subseteq \boldsymbol{A}$ which is the integration of tuples from X being the proper description of the instance.

Procedure:

```
BEGIN
    1. Set $A = \bigcup_{i=1}^{n} T_i$ ;
    2. For each $a \in A$ determine a set with repetitions
           $X_a = \{t_{ia}\text{: } t_i \in X \text{ for } i = 1, 2,\ldots, n\}$;
```

3. For each $a \in A$ using distance function d_a determine a value $v_a \subseteq V_a$ such that

$$\sum_{r_{ia} \in X_a} d_a(v_a, r_{ia}) = \min_{v'_a \subseteq V_a} \sum_{r_{ia} \in X_a} d_a(v'_a, r_{ia})$$

4. Create tuple t^* consisting of values v_a for all $a \in A$;

END.

The most important step in the above algorithm is step 3 which for each attribute determines its integration satisfying criterion C_4. The integration problem of this type has been formulated and analyzed in work [12]. In that work a set of postulates for integration has been defined, and algorithms for its determining have been worked out.

We can prove that the integration determined by this algorithm satisfies all criteria C_1, C_2, C_3 and C_4. The computational complexity of this algorithm is $O(n^2)$.

6 Related Works

The main problem of ontology integration is often formulated as follows: For given ontologies $O_1, \ldots, O_n$ one should determine one ontology which best represents them. In this way we have to deal with ontology merging. For integrating traditional (non-fuzzy) ontologies many works have been done, among others in [4]-[7], [17], [18]. The main contributions of these works are based on methods for matching concepts and their relations. The number of works concerned with integrating ontologies on instance level is not large. Many authors do not treat instances as a part of ontology.

Fuzzy ontology conception is younger and there are not many works on this subject. The same fact is for fuzzy ontology integration. For this kind of ontologies one can distinguish two groups of works. In the first of them the authors proposed logical-based approaches, in which they tried to couple both fuzzy and distributed features using description and fuzzy logics [3], [8]. They worked out several discrete tableau algorithms to achieve reasoning within this new logical system. In the second group the authors proposed a fuzzy ontology generation framework in which a concept descriptor is represented by means of fuzzy relation which encodes the degree of a property value using a fuzzy membership function. The fuzzy ontology integration is based on identifying the most likely location of particular concepts in ontologies to be merged. In [1], [2] the authors proposed an approach to manage imprecise and classic information in databases represented by fuzzy ontology.

Generally in the literature there are missing clear criteria for ontology integration. Most often proposed algorithms refer to concrete ontologies and their justification is rather intuitive than formal. The reason is that the semantics of concepts and their relations are not clearly defined, but rather based on default values.

In this paper we propose to use consensus theory to fuzzy ontology integration on instance level. The advantages of this approach are based on the fact that consensus methods are very useful in processing many kinds of inconsistencies which very often appear in integration tasks. Besides, consensus methods possess well-defined criteria.

As it is well known, ontologies even describing the same real world often contain many inconsistencies because they have been created for different aims. Using consensus methods for integrating fuzzy ontologies is novel since, to the best knowledge of the authors, this approach is missing in the literature. For non-fuzzy ontology integration consensus methods have been used successfully [11], [19].

7 Conclusions

In this paper a method for integrating fuzzy instances is proposed. The structure for fuzzy instance is not based on traditional real number degrees of instance membership to classes, but on the possibility of representing the uncertain and incomplete aspect by using the multi-value structure. Future works should concern implementing the proposed algorithm and verifying the method for real ontologies.

References

1. Abulaish, M., Dey, A.: A Fuzzy Ontology Generation Framework for Handling Uncertainties and Non-uniformity in Domain Knowledge Description. In: Proceedings of the International Conference on Computing: Theory and Applications, pp. 287–293. IEEE (2007)
2. Blanco, I.J., Vila, M.A., Martinez-Cruz, C.: The Use of Ontologies for Representing Database Schemas of Fuzzy Information. International Journal of Intelligent Systems 23(4), 419–445 (2008)
3. Calegari, S., Ciucci, D.: Fuzzy Ontology, Fuzzy Description Logics and Fuzzy-OWL. In: Masulli, F., Mitra, S., Pasi, G. (eds.) WILF 2007. LNCS (LNAI), vol. 4578, pp. 118–126. Springer, Heidelberg (2007)
4. Duong, T.H., Nguyen, N.T., Jo, G.S.: A Method for Integration across Text Corpus and WordNet-based Ontologies. In: IEEE/ACM/WI/IAT 2008 Workshops Proceedings, pp. 1–4. IEEE CS (2008)
5. Duong, T.H., Jo, G.S., Jung, J.J., Nguyen, N.T.: Complexity Analysis of Ontology Integration Methodologies: A Comparative Study. Journal of Universal Computer Science 15(4), 877–897 (2009)
6. Duong, T.H., Nguyen, N.T., Jo, G.S.: A Method for Integrating Multiple Ontologies. Cybernetics and Systems 40(2), 123–145 (2009)
7. Fernadez-Breis, J.T., Martinez-Bejar, R.: A Cooperative Framework for Integrating Ontologies. Int. J. Human-Computer Studies 56, 665–720 (2002)
8. Lu, J., Li, Y., Zhou, B., Kang, D., Zhang, Y.: Distributed Reasoning with Fuzzy Description Logics. In: Shi, Y., van Albada, G.D., Dongarra, J., Sloot, P.M.A. (eds.) ICCS 2007. LNCS, vol. 4487, pp. 196–203. Springer, Heidelberg (2007)
9. Kemeny, J.G.: Mathematics without Numbers. Daedalus 88, 577–591 (1959)
10. Nguyen, N.T.: Using Distance Functions to Solve Representation Choice Problems. Fundamenta Informaticae 48, 295–314 (2001)
11. Nguyen, N.T.: A Method for Ontology Conflict Resolution and Integration on Relation Level. Cybernetics and Systems 38(8), 781–797 (2007)
12. Nguyen, N.T.: Advanced methods for inconsistent knowledge management. Springer, London (2008)

13. Nguyen, N.T.: Consensus system for solving conflicts in distributed systems. Journal of Information Sciences 147, 91–122 (2002)
14. Nguyen, N.T.: Conflicts of Ontologies – Classification and Consensus-Based Methods for Resolving. In: Gabrys, B., Howlett, R.J., Jain, L.C. (eds.) KES 2006, Part II. LNCS (LNAI), vol. 4252, pp. 267–274. Springer, Heidelberg (2006)
15. Nguyen, N.T.: Inconsistency of Knowledge and Collective Intelligence. Cybernetics and Systems 39(6), 542–562 (2008)
16. Nguyen, N.T., Truong, H.B.: A Consensus-Based Method for Fuzzy Ontology Integration. In: Pan, J.-S., Chen, S.-M., Nguyen, N.T. (eds.) ICCCI 2010, Part II. LNCS (LNAI), vol. 6422, pp. 480–489. Springer, Heidelberg (2010)
17. Noy, N.F., Musen, M.A.: SMART: Automated Support for Ontology Merging and Alignment. In: Proc. of the 12th Workshop on Knowledge Acquisition, Modelling and Management (KAW 1999), Banff, Canada, pp. 1–20 (1999)
18. Pinto, H.S., Martins, J.P.: A Methodology for Ontology Integration. In: Proceedings of the First International Conference on Knowledge Capture, pp. 131–138. ACM Press (2001)
19. Stephen, M.L., Hurns, M.N.: Consensus Ontologies: Reconciling the Semantics of Web Pages and Agents. IEEE Internet Computing 5(5), 92–95 (2001)

Making Autonomous Robots Form Lines

Keisuke Satta[1], Munehiro Takimoto[3], and Yasushi Kambayashi[2]

[1] Graduate School of Computer and Information Engineering,
[2] Department of Computer and Information Engineering
Nippon Institute of Technology, 4-1 Gakuendai, Miyashiro-machi,
Minamisaitama-gun, Saitama 345-8501 Japan
c1065241@cstu.nit.ac.jp
yasushi@nit.ac.jp
[3] Department of Information Sciences, Tokyo University of Science
2641 Yamazaki, Noda 278-8510 Japan
mune@cs.is.noda.tus.ac.jp

Abstract. The research and development of various autonomous robots have been conducted for seeking methods of making multiple robots cooperate efficiently. In this paper we report an experiment to control multiple robots by using a set of mobile agents. Our previous study succeeded in making autonomous robots roughly gather by using ACC (Ant Colony Clustering), while suppressing energy consumption. The robots that are gathered are in arbitrary shapes. It is easier for human laborer to collect then if some of them form lines. We have studied to make the robots, which are roughly gathered by using ACC, form short lines. In this paper, we propose the line forming technique of the autonomous robots to achieve the above-mentioned purpose. We have constructed a simulator to show the movements of many robots based on the data collected from a few real robots. The results of the simulation demonstrate the effectiveness of the technique.

Keywords: Autonomous robots, Multiple-robots, Multi-agent system, Intelligent robot control, RFID.

1 Introduction

When we pass through terminals of an airport, we often see carts scattered in the walkway and laborers manually collecting them one by one. It is a laborious task and not a fascinating job. It would be much easier if carts were roughly gathered in any way before the laborers begin to collect them. Multi-robot systems have made rapid progress in various fields, and the core technologies of multi-robots systems are now easily available [5]. Therefore, it is possible to make each cart have minimum intelligence, making each cart an autonomous robot. We realize that for such a system cost is a significant issue and we address one of those costs, the power source. A big powerful battery is heavy and expensive; therefore such intelligent cart systems with small batteries are desirable. Thus energy saving is an important issue in such a system [4]. Travelers pick up carts at designated points and leave them arbitrary

J.-S. Pan, S.-M. Chen, N.T. Nguyen (Eds.): ACIIDS 2012, Part I, LNAI 7196, pp. 198–207, 2012.

places. It is desirable that intelligent carts (intelligent robots) draw themselves together automatically. A simple implementation would be to give each cart a designated assembly point to which it automatically returns when it is free. It is easy to implement, but some carts would have to travel a long way back to their own assembly point, even though they are located close to other assembly points. It consumes too much unnecessary energy. In order to ameliorate the situation, we employ mobile software agents to locate carts scattered in a field, e.g. an airport, and make them autonomously determine their moving behavior using a clustering algorithm based on ant colony optimization (ACO). ACO is a swarm intelligence-based method and a multi-agent system that exploits artificial stigmergy for the solution of combinatorial optimization problems. Preliminary experiments yield a favorable result. Ant colony clustering (ACC) is an ACO specialized for clustering objects. The idea is inspired by the collective behaviors of ants, and Deneubourg formulated an algorithm that simulates the ant corps gathering and brood sorting behaviors [1].

We previously proposed the ACC approach using mobile software agents. Quasi-optimal carts collection is achieved through the repetitions of the three phases. The first phase collects the positions of carts. One mobile agent issued from the host computer visits scattered carts one by one and collects the positions of them. The precise coordinates and orientation of each cart are determined by sensing RFID (Radio Frequency Identification) tags under the floor carpet. Upon the return of the position collecting agent, the second phase begins wherein another agent, the simulation agent, performs the ACC algorithm and produces the quasi-optimal gathering positions for the carts. The simulation agent is a static agent that resides in the host computer. In the third phase, a number of mobile agents are issued from the host computer. Each mobile agent migrates to a designated cart, and drives the cart to the assigned quasi-optimal position that is calculated in the second phase. One step of the collection task consists of these three phases, and each step contributes the collection. After many steps, the collection task should be achieved. Once the quasi-optimal gathering positions are calculated, the simulator performs another simulation that imitates all the behaviors of all the carts, and produces the procedure that each cart follows. The procedure has each route and waiting timing for each cart should trace. Since the precise route and waiting timing for each cart is calculated, the driving agent on each cart has all the knowledge to drive its cart to the tentative goal point.

As described above, the cart robot was able to roughly gather the scattered in a field. Even though cart collecting clustering methods have been intensively researched, simple clusters of carts are not very helpful for manual laborers [7] [8]. They not only collect carts but also they have to arrange them in lines. For implementing a practical system, lining up collected carts is essential. Therefore, in this paper we discuss how to make collected cart robots be arranged into chunks of short lines. Through experiments, both in real robots and in a simulator, we demonstrate the method we propose is useful for the autonomous robots forming lines.

The structure of the balance of this paper is as follows. In the second section, we describe the background. In the third section, we describe the agent system that performs the arrangement of the intelligent carts. The agent system consists of several static and mobile agents. The static agents interact with the users and compute the

ACC algorithm, and other mobile agents gather the initial positions of the carts and distribute the assembly positions. The fourth section describes how the collected cart robots arrange themselves into lines. Through experiments, both in real robots and in a simulator, we demonstrate the method we propose is useful for the autonomous robots forming lines. The fifth section describes the simulator and demonstrates the feasibility of our intelligent cart gathering system. In the section, we report the results we have obtained from the experiments on the simulator. Finally, in the sixth section we summarize the work and discuss future research directions.

2 Background

Kambayashi and Takimoto have proposed a framework for controlling intelligent multiple robots using higher-order mobile agents [4][5]. The framework helps users to construct intelligent robot control software by migration of mobile agents. Since the migrating agents are higher-order, the control software can be hierarchically assembled while they are running. Dynamically extending control software by the migration of mobile agents enables them to make base control software relatively simple, and to add functionalities one by one as they know the working environment. Thus they do not have to make the intelligent robot smart from the beginning or make the robot learn by itself. They can send intelligence later as new agents. Even though they demonstrate the usefulness of the dynamic extension of the robot control software by using the higher-order mobile agents, such higher-order property is not necessary in our setting. They have implemented a team of cooperative search robots to show the effectiveness of their framework, and demonstrated that their framework contributes to energy saving of multiple robots [4]. They have achieved significant saving of energy.

Deneubourg has formulated the biology inspired behavioral algorithm that simulates the ant corps gathering and brood sorting behaviors [1]. Lumer has improved Deneubourg's model and proposed a new simulation model that is called Ant Colony Clustering [6]. His method could cluster similar objects into a few groups. Kambayashi et al have improved Lumer's model and proposed a multi-robot control by using ACC [2]. This method does not cluster similar objects but cluster nearby objects by using pheromone. Kambayashi et al have designed an intelligent cart system based on the proposed method [3]. As mentioned above, previous studies could collect mobile robots roughly using ACC. In this paper, we present a technique to form a series of short lines after collecting mobile robots.

3 System Model

Our system model consists of carts and a few kinds of static and mobile software agents. All the controls for the mobile carts as well as ACC computation performed in the host computer are achieved through the static and mobile agents. They are: 1) user interface agent (UIA), 2) operation agents (OA), 3) position collecting agent (PCA), 4) clustering simulation agent (CSA), and 5) driving agents (DA). All the software

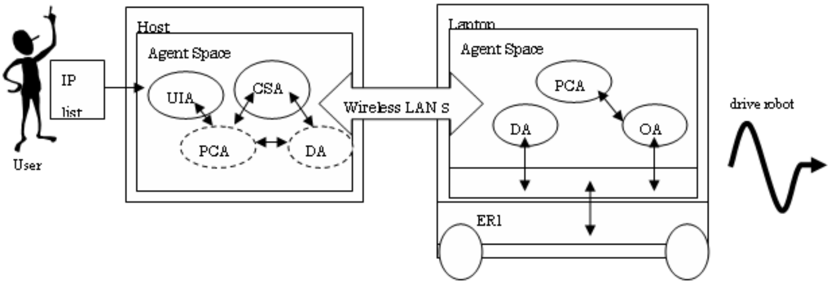


Fig. 1. Cooperative agents to control an intelligent cart

agents except UIA and CSA are mobile agents. A mobile agent traverses carts scattered in the field one by one to collect their coordinates. After receiving the assembly positions computed by a static agent, many mobile agents migrate to the carts and drive them to the assembly positions. Fig. 1 shows the interactions of the cooperative agents to control an intelligent cart. The followings are details of each agent:

1) User Interface Agent (UIA): The user interface agent (UIA) is a static agent that resides on the host computer and interacts with the user. It is expected to coordinate the entire agent system. When the user creates this agent with a list of IP addresses of the intelligent carts, UIA creates PCA and passes the list to it.
2) Operation Agent (OA): Each cart has at least one operation agent (OA). It has the task that the cart on which it resides is supposed to perform. Each intelligent cart has its own OA. Currently all operation agents (OA) have a function for collision avoidance and a function to sense RFID tags embedded in the floor carpet to detect its precise coordinates in the field.
3) Position Collecting Agent (PCA): A distinct agent called position collecting agent (PCA) traverses carts scattered in the field one by one and collects their coordinates. PCA is created and dispatched by UIA. Upon returning to the host computer, it hands the collected coordinates to the clustering simulation agent (CSA) for ACC.
4) Clustering Simulation Agent (CSA): The host computer houses the static clustering simulation agent (CSA). This agent actually performs the ACC algorithm by using the coordinates collected by PCA as the initial positions, and produces the quasi-optimal assembly positions of the carts, and then performs yet another simulation to produce instructions each cart follows to reach its assigned goal position. Upon terminating the simulation and producing the procedures for all the intelligent carts, CSA creates a number of driving agents (DA).
5) Driving Agent (DA): The quasi-optimal arrangement coordinates, as well as procedures to reach them, produced by the CSA are delivered by driving agents (DA). One driving agent is created for each intelligent cart, and it contains the set of procedures for the cart. The DA drives its intelligent cart to the designated assembly position. DA is the intelligent part of the intelligent cart.

OA detects the current coordinate of the cart on which it resides. Each cart has its own IP address and UIA hands in the list of the IP addresses to PCA. First, PCA migrates to an arbitrary cart and starts hopping between them one by one. It communicates locally with OA, and writes the coordinates of the cart into its own local data area. When PCA gets all the coordinates of the carts, it returns to host computer. Upon returning to the host computer, PCA creates CSA and hands in the coordinate data to CSA which computes the ACC algorithm.

The current implementation employs RFID (Radio Frequency Identification) to get precise coordinates [9]. We set RFID tags in a regular grid shape under the floor carpet tiles. The tags we chose have a small range so that the position-collecting agent can obtain fairly precise coordinates from the tags. Also, the cart has a basic collision avoidance mechanism using infrared sensors.

CSA is the other static agent and its sole role is ACC computation. When CSA receives the coordinate data of all the carts, it translates the coordinates into coordinates for simulation, and then performs the clustering. When CSA finishes the computation and produces a set of assembly positions, it then creates the set of procedures for autonomous cart movements.

Then CSA creates DA that conveys the set of procedures to the intelligent carts as many. Each DA receives its destination IP address from PCA, and the set of procedures for the destination cart, and then migrates to the destination cart. Each DA has a set of driving procedures that drives its assigned cart to the destination, while it avoids collision. OA has the basic collision detection and avoidance procedures, and DA has task-specific collision avoidance procedures.

4 Forming Lines

As shown in Fig. 2, we have successfully collected robots at quasi-optimal positions from scattered robots in the field using ACC. Simple clusters of robots are not very helpful for practical purpose. We need to arrange them in order by some ways.

In this section, we discuss how to make collected cart robots be arranged into chunks of short lines. The Alignment method we are proposing consists of two major phases. In order to achieve line-forming, we extend DA so that they can interact with DAs in nearby neighbor robots. In the first phase, nearby robots with similar directions arrange themselves in short lines. In the second phase, isolated robot attempt to adhere to already formed lined robots. This phase should make all the robots form into groups. We describe the two phases in detail below.

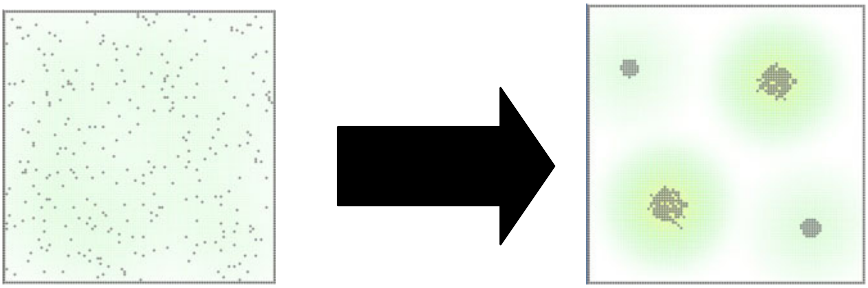


Fig. 2. The simulated gathering positions Fig.

1. The First Phase

In the first phase, each of two nearby robots get together to form a short lined cluster as shown in Fig. 3. They become the core of serialized clusters. This is also done through agent migrations. In order to achieve this migration, we extended DA's functionalities as well as we added small cheap web camera to each robot as follows. Each robot has a color board so that when another robot happens to be nearby and faces similar direction, it can recognize the board. The web camera we set is cheap and short sighted, but is appropriate enough for our purpose.

(1) When a robot A finds robot B through Web camera, it means robot B is close and has similar direction, and the DA on the robot A create a clone of it and send it to robot B.

(2) When the dispatched agent arrives at the robot B, it overrides the DA on robot B and start to drive it. The dispatched agent drives the robot B to adjust the location and angle so that the robot B is exactly on the line the same as the robot A's direction and it faces the same direction as the robot A, so that the robot A can adhere to the robot B as it simply forwards. When the adjustment is done the agent moves back to the robot A to inform it is ready.

(3) Robot A forwards to the back of B, and adheres to it.

If a robot cannot find any colleagues nearby, it just remains in the place without changing the direction or the position.

2. The Second Phase

The second phase tries to sweep the isolated robots and put them into some lines formed in the first phase. Between the first phase and the second phase, the PCA itinerates once again, and collects all the positions of the robots so that each isolated robot can know where the closest formed line is. Upon understanding the position of the line it is supposed to adhere, the DA on an isolated robot drives its robot toward the position, and when it comes close enough to it, it repeats the same behavior as the first phase as shown in Fig. 4. Here robot A is the isolated robot.

(1) The DA on the robot A drives the robot to adjust the location and angle so that the robot B is exactly on the line the same as the robot B's direction and it faces the same direction as the robot B, so that the robot B can adhere to the robot A as it simply forwards. When the adjustment is done the agent create a clone of it and sends to the robot B to inform it is ready.

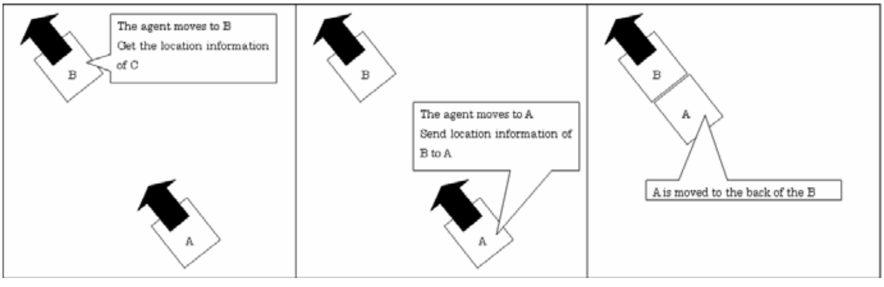

The agent moves to B
Get the location information of C
B
A
B
The agent moves to A
Send location information of B to A
A
B
A
A is moved to the back of the B

Fig. 3. The first behavior

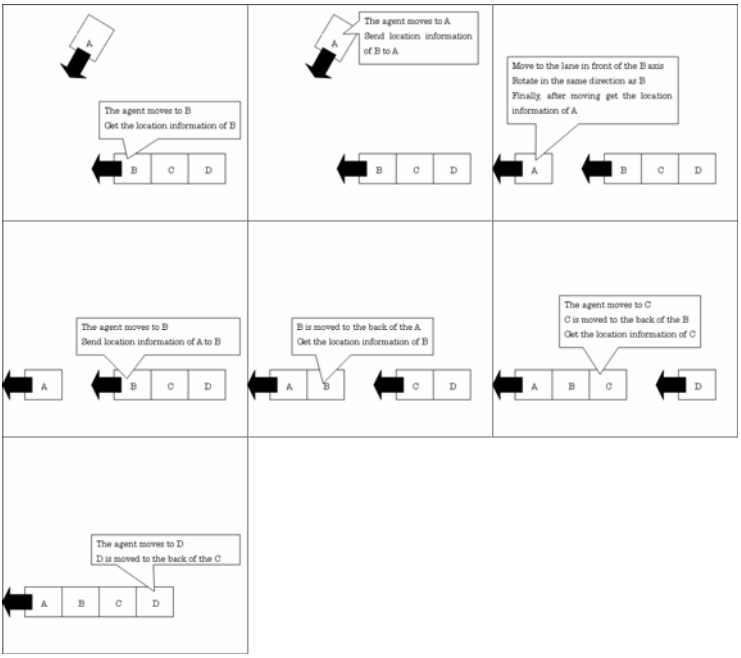

A
The agent moves to B
Get the location information of B
B
C
D
A
The agent moves to A
Send location information of B to A
B
C
D
Move to the lane in front of the B axis
Rotate in the same direction as B
Finally, after moving get the location information of A
A
B
C
D
A
The agent moves to B
Send location information of A to B
B
C
D
B is moved to the back of the A
Get the location information of B
A
B
C
D
The agent moves to C
C is moved to the back of the B
Get the location information of C
A
B
C
D
The agent moves to D
D is moved to the back of the C
A
B
C
D

Fig. 4. The second behavior

(2) When the cloned agent arrives at the robot B, it overrides the DA on robot B and start to drive it. This time, a combination of location information as well as information from the web camera is used to achieve the task as is in the first phase. Since the cloned agent comes from the robot A, it knows the color of the board of the robot A and forwarding to the robot and makes it adhere to the robot A is straightforward.

(3) Since robot B is the head of a lined robots, e.g. robots B, C, and D form a line as a group, we need to make all of them go forward. This is done through the dispatched agent's migration as shown in Fig. 4. For example, in Fig. 4, after the agent drives the robot B to the back of the robot A, it migrate to the robot C, and then robot D. When it finds there are no more robots to drive, it simply kills itself.

Fig. 5 shows the experiment for the first phase by using two real robots.

5 Experiments

We have conducted two kinds of experiments to demonstrate our framework for the intelligent robot carts is feasible. The first one is to check whether two relatively close robots can form a line. As shown in Fig. 5, this is achieved successfully. Since the task is mainly depends of the web camera. They need to be real close. But it should be alright. Considering the cart application, the system should require cheap web cameras. Isolated remaining robot can be assembled in the second phase.

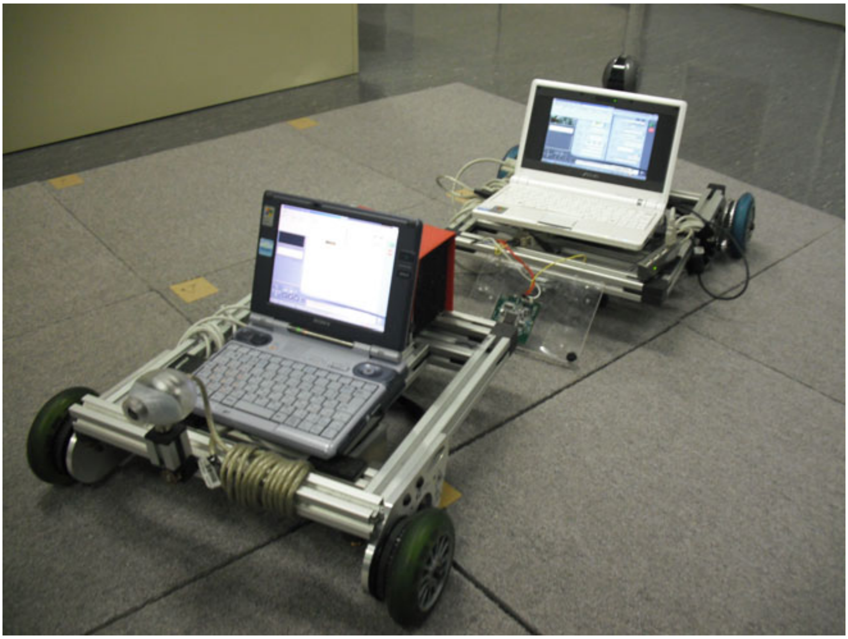


Fig. 5. The actual alignment

For the second phase, we have built a simulator. As shown in Fig. 6, the field is 50 × 50 square grid. The number placed of robots is 50. The robots' initial positions are randomly determined. As shown in Fig. 6, quit a few of the robots get together to form short lines. Even though they are far from the perfect, forming several short lines is one big leap from randomly assembled clusters in the previous work.

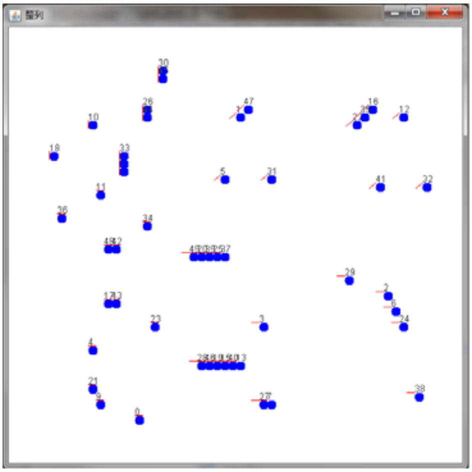


Fig. 6. Simulation behavior results

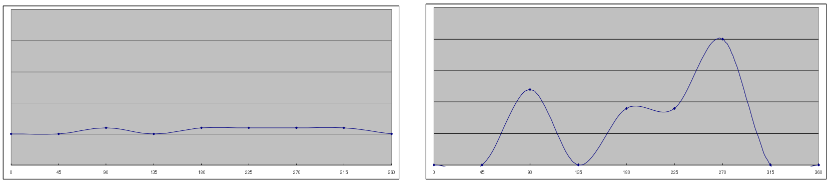


Fig. 7. Scatter plot of the direction in which alignment before and after alignment

In order to numerically evaluate our achievement, we have measured the variance. Fig. 7 shows a scatter plot of the direction of robots before and after alignment. As the before-alignment figure (left hand side) shows all the robots are facing completely random. It shows even distribution. On the other hand, as the after-alignment figure (right hand side) shows there are some strong biases. We can conclude that certain number of robots successfully form lines. We believe that more elaborate algorithm certainly sorts robots more clearly, and we are certain we are on the right track.

6 Summary

This paper shows a framework for controlling the robots, which are connected to a communication network, to move into groups. The framework for the ACC to the mobile robot in scattered field will be set to an optimal position on the field, followed by fine alignment is performed. Through experiments we have demonstrated that our algorithm makes the robots assembled into some short lines. Considering cart application, making scattered robots into several short lines is a big leap toward a practical system. Shintani et al conduct similar research project [10]. They mainly use good web camera to get coordinates of other robots, and two kinds of mobile agents namely pheromone agents and driving agent. They actually achieve more elaborate system to serialize robots, but use more expensive equipment. Our approach, i.e. using cheap equipment, we believe more practical. The experiments we have conducted for a large number of robots are through simulator not actual robots. The simulator, however, shows good results, and actually we have conducted another project that is building real intelligent shopping carts that follow the users [11]. We can combine those achievements into one big project so that we can build a practical intelligent cart robot for assisting both the users and the collecting laborers.

References

1. Deneubourg, J., Goss, S., Franks, N., Sendova-Franks, A., Detrain, C., Chretien, L.: The Dynamics of Collective Sorting: Robot-Like Ant and Ant-Like Robot. In: First Conference on Simulation of Adaptive Behavior: From Animals to Animats, pp. 356–363. MIT Press (1991)
2. Kambayashi, Y., Ugajin, M., Sato, O., Tsujimura, Y., Yamachi, H., Takimoto, M., Yamamoto, H.: Integrating Ant Colony Clustering Method to a Multi-Robot System Using Mobile Agents. Industrial Engineering and Management System 8(3), 181–193 (2009)

3. Kambayashi, Y., Ugajin, M., Sato, O., Takimoto, M.: Design of an Intelligent Cart System for Common Airports. In: 13th IEEE International Symposium Consumer Electronics, CD-ROM (2009)
4. Takimoto, M., Mizuno, M., Kurio, M., Kambayashi, Y.: Saving Energy Consumption of Multi-robots using Higher-Order Mobile Agents. In: Nguyen, N.T., Grzech, A., Howlett, R.J., Jain, L.C. (eds.) KES-AMSTA 2007. LNCS (LNAI), vol. 4496, pp. 549–558. Springer, Heidelberg (2007)
5. Kambayashi, Y., Takimoto, M.: Higher-order mobile agents for controlling intelligent robots. International Journal of Intelligent Information Technologies 1(2), 28–42 (2005)
6. Lumer, E.D., Faiesta, B.: Diversity and adaptation in populations of clustering ants, from animals to animats 3. In: 3rd International Conference on the Simulation of Adaptive Behavior, pp. 501–508. MIT Press, Cambridge (1994)
7. Kambayashi, Y., Yamachi, H., Takimoto, M.: A Feasibility Study of the Intelligent Cart System. In: SICE Annual Conference, pp. 1159–1163 (2010)
8. Kambayashi, Y., Yamachi, H., Takimoto, M.: A Search for Practical Implementation of the Intelligent Cart System. In: International Congress on Computer Applications and Computational Science, pp. 895–898 (2010)
9. Kambayashi, Y., Takimoto, M.: Location of Intelligent Carts Using RFID. In: Turcu, C. (ed.) Deploying RFID: Challenges, Solutions, and Open Issues, pp. 249–264. InTech, Rijeka (2011)
10. Shintani, M., Lee, S., Takimoto, M., Kambayashi, Y.: A Serialization Algorithm for Mobile Robots Using Mobile Agents with Distributed Ant Colony Clustering. In: König, A., Dengel, A., Hinkelmann, K., Kise, K., Howlett, R.J., Jain, L.C. (eds.) KES 2011, Part I. LNCS (LNAI), vol. 6881, pp. 260–270. Springer, Heidelberg (2011)
11. Kohtsuka, T., Onozato, T., Tamura, H., Katayama, S., Kambayashi, Y.: Design of a Control System for Robot Shopping Carts. In: König, A., Dengel, A., Hinkelmann, K., Kise, K., Howlett, R.J., Jain, L.C. (eds.) KES 2011, Part I. LNCS (LNAI), vol. 6881, pp. 280–288. Springer, Heidelberg (2011)

Genetic Algorithm-Based Charging Task Scheduler for Electric Vehicles in Smart Transportation*

Junghoon Lee[1], Hye-Jin Kim[1], Gyung-Leen Park[1,**], and Hongbeom Jeon[2]

[1] Dept. of Computer Science and Statistics, Jeju National University
[2] Smart Green Development Center, KT, Republic of Korea
{jhlee,hjkim82,glpark}@jejunu.ac.kr, hbjeon@kt.com

Abstract. This paper presents a design and evaluates the performance of an efficient charging scheduler for electric vehicles, aiming at reducing the peak load of a fast charging station while meeting the time constraint of all charging requests. Upon the task model consist of actuation time, operation length, deadline, and a consumption profile, the proposed scheduler fills the allocation table, by which the power controller turns on or off the electric connection switch to the vehicle on each time slot boundary. For the sake of combining the time-efficiency of heuristic-based approaches and the iterative evolution of genetic algorithms, the initial population is decided by a heuristic which selects necessary time slots having the lowest power load until the previous task allocation. Then, the regular genetic operations further improve the schedule, additionally creating a new chromosome only from the valid range. The performance measurement result obtained from a prototype implementation shows that our scheme can reduce the peak load for the given charging task sets by up to 4.9 %, compared with conventional schemes.

Keywords: smart grid, electric vehicle charging, genetic algorithm, initial population, peak load reduction.

1 Introduction

Built on top of advanced information and two-way digital communication technologies, the smart grid cannot just intelligently deliver electricity but also manage power facilities taking advantage of real-time information exchange and interaction between providers and consumers [1]. Meanwhile, electric vehicles, or EVs in short, are new and important targets for the smart grid to manage, as they require the use of batteries having high energy-storage capacity as well as bringing large electric load for their charging [2]. EVs are indispensably forced to be equipped with an electronic interface for grid connection to allow controlled

* This research was supported by the MKE, Korea, under IT/SW Creative research program supervised by the NIPA (NIPA-2011-(C1820-1101-0002)).

** Corresponding author.

J.-S. Pan, S.-M. Chen, N.T. Nguyen (Eds.): ACIIDS 2012, Part I, LNAI 7196, pp. 208–217, 2012.

energy exchanges. Even though many researches are trying to improve the driving range while cutting down the charging time, EVs still need to be charged more often and it takes at least tens of minutes [3]. Accordingly, it is necessary to build a nationwide charging infrastructure which embraces fast charging stations, battery swapping stations, and individual charging points for slower charging [2].

For the large scale deployment of EVs, the smart grid can have a centralized management architecture and also local control entities working in administrative units such as a building or a charging station. EV penetration will put increased pressure on peaking unit not having any charging control strategy [4]. Moreover, some grids are likely to build additional generation capacity to catch up with the increased power demand resulted from concentrated EV charging. To cope with this problem, a grid unit can run demand management programs, which are becoming more important to meet the customer requirement as well as to achieve system goals like peak load reduction, power cost saving, and energy efficiency. Demand response schemes can shed or shift peak load according to the load type, mainly by scheduling the given charging tasks [1]. Moreover, via the global network connection, they can further consider price signal change and current load condition.

The scheduling problem for charging tasks is quite similar to process scheduling in the real-time operating system, as each charging operation can be modeled as tasks having execution time, start time, and a deadline [5]. The difference lies in that charging tasks can run in parallel as long as the total power does not exceed the provisioned capacity. Task scheduling is in most cases a complex time-consuming problem sensitive to the number of tasks. It is difficult to solve by conventional optimization schemes, and their severe execution time makes them impractical in the real system. For such a problem, genetic algorithms can provide efficient search techniques based on principles of natural selection and genetics. Specifically, the convergence and performance of them are inherently affected by the initial generation, especially when the solution space is large and complex [6]. If the initial population has better quality, the genetic algorithm is more likely to yield a better result.

Moreover, heuristic-based approaches also gain computational performance possibly at the cost of accuracy, mainly taking advantage of empirical intelligence. Provided that the heuristic is fast and efficient enough, it is possible to make their solutions included in the initial population and run a genetic algorithm to further improve the quality of feasible solutions. This strategy can combine the time-efficiency of heuristic-based approaches and the iterative evolutions of genetic algorithms. Based on this motivation, this paper designs an efficient charging task scheduler for EVs, aiming at reducing the peak load in a charging station and thus remedying the problem of power demand concentration in a specific time interval. Here, a fast heuristic is available by our previous work [7], which finds reasonable solutions with the polynomial time complexity for the number of charging tasks.

This paper is organized as follows: After issuing the problem in Section 1, Section 2 introduces the background and related work. Section 3 explains the target system architecture and designs a charging task scheduler. After performance measurement results are demonstrated and discussed in Section 4, Section 5 summarizes and concludes the paper with a brief introduction of future work.

2 Background and Related Work

As the large deployment of EVs is commonly expected for energy efficiency and pollution reduction, the management of a power load induced by their concentrated charging is a major concern in the smart grid [8]. In this aspect, [2] presents a conceptual framework to successfully integrate EVs into electric power systems. Focusing on slow charging, the framework covers the grid technical operation and the electricity market environment, while their simulation study discovers that the importance of unit-level power control. In addition, [4] deals with the development of a demand response model for residential customers facing EV penetration to study the impact of customer responses to different pricing policies on the distribution-level load shapes. The analysis is conducted based on charging profiles, charging strategies, and EV penetration scenarios. It addresses that the time-of-use rate can be properly designed to reduce the peak demand for the large deployment of EVs.

In genetic algorithms, the initial population is sometimes seeded with some of those feasible solutions or partial solutions of the problem. However, the initial population selection itself is a complete algorithm or heuristic for the given optimization problem. [9] takes into account search space, fitness function, diversity, problem difficulty, selection pressure, and the number of individuals in deciding the initial population. The authors assert that the more diverse the initial population is, the greater the possibility to find a better solution. The diversity is measured 1) at the gene level via the Grefenstette bias formula and uncertainty entropy, 2) at the chromosome level via hamming distance and neighborhoodness, and 3) at the population level via the center of mess, respectively. This scheme suggests a metric in randomly generated genes within a population. It is expected that the computation time needed in calculating diversity and the initial population can be rewarded by the quality of solution, reduced number of iterations, and the like.

3 Charging Scheduler Design

3.1 System Model

In our service scenario shown in Figure 1, a driver tries to make a reservation at a charging station via a vehicular network specifying its requirement details, while he or she is driving. Each requirement consists of vehicle type, estimated arrival time, the desired service completion time (deadline), charging amount, and so on. A request can be issued also via the on-site connection to the charging facility. Receiving the request, the scheduler prepares the power load profile of the vehicle type from the well-known vehicle information database. The load profile, or interchangeably consumption profile, contains the power consumption dynamics along the time axis for specific EV charging. Then, the station checks whether it can meet the requirement of the new request without violating the constraints of already admitted requests. The result is delivered back to the vehicle, and the driver may accept the schedule, attempt a renegotiation, or choose another station. The accuracy of load profile is critical to the correct schedule generation, but we assume that sufficiently accurate profile is available to focus on the scheduling scheme [4].

From the vehicle-side viewpoint, entering the station, the vehicle is assigned and plugged to a charger as shown in Figure 1. The controller connects or disconnects power to each vehicle according to the schedule generated by either a scheduler within a charging station or a remote charging server running in the Internet. Nowadays, a high-capacity server in the cloud can provide computing power to time-intensive applications. The SAE J1772 series define the standard for electric connectors and their charging system architecture, including physical, electrical, communication protocol, and performance requirement [10]. Our scheduler can work with this standard as a core service for electric vehicles. It must be mentioned that the interaction between the scheduler and EVs can be performed through the specific vehicle network, such as cellular networks, vehicle ad hoc networks, or a combination of them under the control of vehicle telematics system.

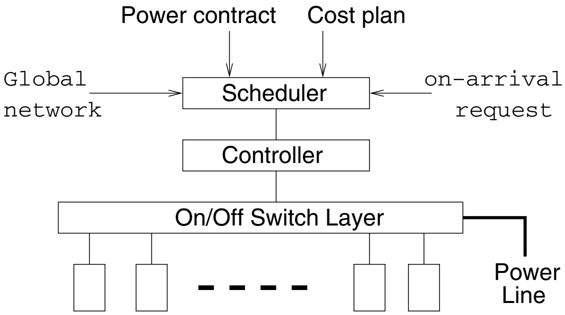


Fig. 1. Charging station model

3.2 Task Model

Each charging operation can be modeled as a task. Task T_i can be described with the tuple of $< A_i, D_i, U_i >$. A_i is the activation time of T_i, D_i is the deadline, and U_i denotes the operation length, which corresponds to the length of the load profile entry. A_i is the estimated arrival time of the vehicle including small margin. It's true that the estimation cannot be always accurate, but modern telematics technologies reduce the prediction error through current traffic information and an efficient path finding algorithm. If an EV arrives earlier than A_i, it can wait with its charging interface connected to the jack. Here, we assume that available jacks and waiting space are enough to accommodate all arriving vehicles. Even if an EV doesn't arrive within A_i, the scheduler can generate a new allocation promptly if the computation time is not severe.

For a task, the power consumption behavior can vary according to the charging stage, remaining amount, vehicle type, and the like. The power consumption profile is practical for characterizing the consumption dynamics along the battery charging stage. This profile is the basic information in generating a charging schedule. In the profile, the power demand is aligned to the fixed-size time slot, during which the power consumption is constant considering the availability of automatic voltage regulation [11]. The slot length can be also affected by power charging dynamics. The length of a time slot can be tuned according to the system requirement on the schedule granularity and the computing time, and likely coincides with the time unit generally used in the real-time price signal. The charging operation can start only at the slot boundary for scheduling efficiency.

3.3 Initial Population Selection

The scheduling process is to fill an $M \times N$ allocation table, where M is the number of slots and N is the number of EVs to charge. The size of M, namely, the scheduling window, depends on the charging station policy on advance booking. The policy selects the value of M so as to cover many task deadlines. Anyway, we set M to 20, and if the slot size is 10 minutes, the scheduling window will be 200 minutes. The allocation procedure fills the allocation table from the first row, each row being associated with a task. Basically, the procedure can create the search space for all feasible allocations. For a single charging task, there can be ${}_{(D_i - A_i)}C_{U_i}$ allocations in its row. It imposes an intolerable computation delay when M or N gets large. To overcome this problem, our basic strategy looks into the allocation up to T_{i-1} and takes U_i slots having the smallest assigned power out of $D_i - A_i$ slots for T_i [12]. Considering $O(1)$ space complexity, we have a lot of margin to further investigate another allocation.

However, the basic allocation is sensitive to the allocation order for the given set, namely, which task to allocate first. In addition, especially for the first task, all slots are not assigned yet. It is impossible to decide which one is better. The allocation procedure would assign slots without intermission if no other restriction is given. It narrows selectable options for the next tasks. To give more flexibility to the subsequent task allocation, our procedure randomly selects for tie-breaks. According to how many first tasks to randomly allocate, there can be more than one allocation for a single order. Even though this procedure increases the search space complexity several times but it still remains in $O(1)$ with time complexity of $O(N^2)$ added. Now, we consider the allocation order according to deadline, length, and per-slot power requirement. That is, our scheme sorts the given task set according to those criteria, applies the basic strategy to each of the corresponding orders, and finds the best schedule among them.

First, the slack is defined as the difference between the deadline and the last start time to meet the time constraint. The larger the slack, the task has more options in its schedule generation. As for ordering by slack, tasks having fewer options are placed first. Then, those tasks having relatively more options are more likely to find slots having less power consumption. Next, the task having longer operation length fills more table entries. If smaller-length tasks are allocated first, the longer tasks will smooth the peak. Last, we can order the task according to the weight, or per-slot power demand. The weight for a task is the average power demand during its operation time. If tasks demanding more power meet at the same time slot, the peak will get too large. To avoid this situation, our scheme allocates those tasks first, as the next allocation can distribute the power-intensive slots. While [7] selects the best one out of those allocations, this paper makes all of them included in the initial population of a genetic algorithm. The rest of population is filled randomly as usual.

3.4 Genetic Operations

In genetic algorithms, with the initial population consist of feasible solutions, they iteratively run genetic operations such as reproduction, mutation, and crossover to improve the quality of solutions. Each evolutionary step generates a population of candidate solutions and evaluates the population according to a fitness function to select the best solution and mate to form the next generation. Over a number of generations, good traits dominate the population, resulting in an increase in the quality of solutions. It must be mentioned that the genetic algorithm process can possibly run for years and does not find any better solution than it did in the first part of the process.

In the charging task scheduling problem, a chromosome corresponds to a single feasible schedule, and is represented by a fixed-length string of an integer-valued vector. The charging task can be started, suspended, and resumed at a slot boundary. Each value element indicates an allocation map for a task. For T_i, U_i slots must be selected out of slots from A_i to D_i. Suppose that $D_i - A_i$ is 5 and U_i is 3. This is the same example as in our previous work [13]. There can be ${}_5C_3$ feasible allocation maps for this task from (0, 0, 1, 1, 1) to (1, 1, 1, 0, 0), where 1 means that charging is performed at the corresponding slot. $1 \rightarrow 0$ transition denotes suspension, while $0 \rightarrow 1$ transition denotes resumption of charging. The decimal equivalent of this binary map will be a vector element of a chromosome. If the vector element is 11, namely, its map is (0, 1, 0, 1, 1), the profile entry of the task is (2, 3, 4), and A_i is 15, it will be mapped to (0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 2, 0, 3, 4). Likewise, an allocation vector can be converted into the corresponding allocation table. For each allocation, the scheduler can calculate the per-slot power requirement and the peak load to evaluate fitness.

The iteration consists of selection and reproduction. Selection is a method that picks parents according to the fitness function. The Roulette Wheel selection gives more chances to chromosomes having better fitness values for mating. Reproduction, or crossover, is the process taking two parents and producing a child with the hope that the child will be a better solution. This operation randomly selects a pair of two crossover points and swaps the substrings from each parent. Reproduction may generate a same chromosome as the existing ones in the population. It is meaningless to have multiple instances of a single schedule. So, they will be replaced by new random ones. Additionally, mutation exchanges two elements in a chromosome. However, each element has a different permissible range, so the mutation must be prohibited. The charging scheduler is subject to time constraint. But, this constraint can be always met, as the scheduler selects the allocation vector only within the valid range.

4 Performance Measurement

This section implements the proposed allocation method using Visual C++ 6.0, making it run on the platform equipped with Intel Core2 Duo CPU, 3.0 GB memory, and Windows Vista operating system. M is set to 20, so if a single time unit is 10 min, the total scheduling window will be 200 min. For a task, the start time is selected randomly between 0 and $M - 1$, while the operation length exponentially distributes. A task will be discarded and replaced if the finish time, namely, the sum of start time and the operation length, exceeds M. In addition, the power level for each time slot ranges from 1 through 5. The power scale is not explicitly specified in this experiment, as it is a relative-value term. For each parameter setting, 50 tasks are generated, and the results are averaged.

For performance comparison, we select 3 policies. First, the *Smallest* allocation is our heuristic [7], which is actually the best of the initial population in the proposed scheme. Second, the *Genetic* allocation runs general genetic operations iteratively with the initial population randomly set. Third, the *Random* allocation generates feasible schedules using the random numbers during the approximately same time interval needed to execute the proposed scheme. It randomly selects U_i out $D_i - A_i$ slots. The *Random* selection works quite well as the candidate allocation is selected only within the valid range for each set. Particularly, if the difference between U_i and $D_i - A_i$ is small for some tasks and the number of tasks is small, this scheme can be sometimes efficient. After all, for fair comparison, the same task set is given to the 4 schemes in each parameter setting.

The first experiment measures the peak load reduction according to the number of tasks. In *Genetic* and proposed schemes, the population size is set to 60 and the number of iterations is set to 1,000. The operation length and the slack of a task exponentially distribute with the average of 5.0 and 2.0, respectively, while the number of tasks ranges from 5 to 20. Figure 2 plots the peak load for each allocation scheme. *Random*, *Smallest*, and *Genetic* schemes show almost same performance, while the proposed scheme reduces peak load by up to 4.9 % for 20 task case. When there are less than 10 tasks, the *Random* scheme yields a little bit better result than *Genetic* and *Smallest* schemes, as it can try a lot of allocations benefiting from its simplicity. However, beyond 10 tasks, the others are better.

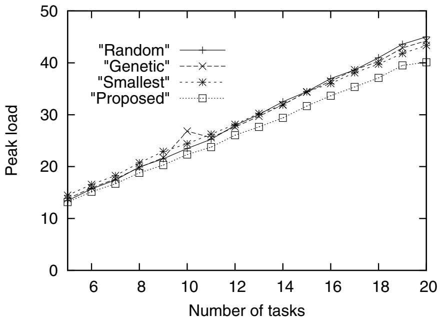


Fig. 2. Effect of the number of tasks

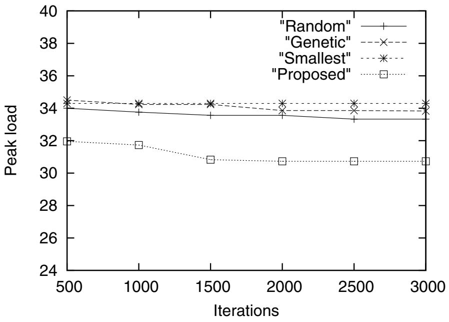


Fig. 3. Effect of iterations

The next experiment measures the effect of the number of iterations and the result is exhibited in Figure 3. Here, other parameters are the same as the previous experiment, but the number of tasks is fixed to 15. The *Smallest* scheme is not affected by the number of iterations. This scheme is uncomparably fast, but it shows almost the same performance as *Genetic* and *Random* schemes. The proposed scheme, basically outperforming others by about 3.5 %, can further reduce the peak load by 1.5 % by the extended iterations, namely from 500 to 3,000, while the *Genetic* scheme by 0.5% and the *Random* scheme by 0.3%. In our observation, most task sets reaches stable peak loads before 200 iterations, and are scarcely further improved.

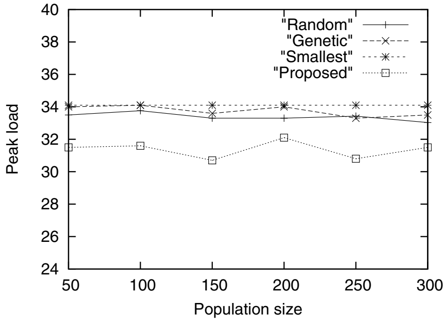


Fig. 4. Effect of population size

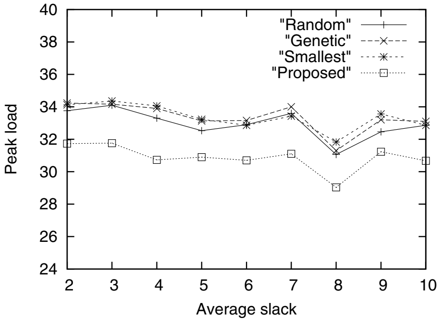


Fig. 5. Effect of slack

In addition, Figure 4 plots the effect of population size to the peak load. The *Smallest* scheme shows the poorest performance, while our scheme improves the peak load by about 3.4 % over the whole range. The peak load largely decreases according to the increase of the population size except when it is 100. It can result from an extraordinary peak load in some task sets. Anyway, Figure 4 finds out that the effect of population size is not so significant in vehicle charging schedule. Moreover, Figure 5 shows the peak load when the slack changes from 2.0 to 5.0

slots. The larger the slack, the more flexible schedule we can get. For the given slack range, our scheme gets 2.6 % reduction in the peak load, while the others get 3.0 %, 2.9 %, and 2.5 %. 4 schemes show almost the same peak load change pattern.

Finally, how many ordered allocations to insert to the initial population, namely, combination ratio, is an additional performance parameter. Figure 6 plots the result. When the combination ratio is 0, the proposed scheme is identical to the *Genetic* scheme, as the initial population is selected purely randomly. Additionally, the *Smallest* scheme will be the same as the uncoordinated allocation, which allocates the operation as soon as the task is ready without permitting preemption. Except when the combination ratio is 0, the proposed scheme does not change much, just slightly reducing the peak load. For all range, the proposed scheme outperforms others by about 2.6 %. The chromosome not the best in the heuristic can contribute to improving the quality of next generations.

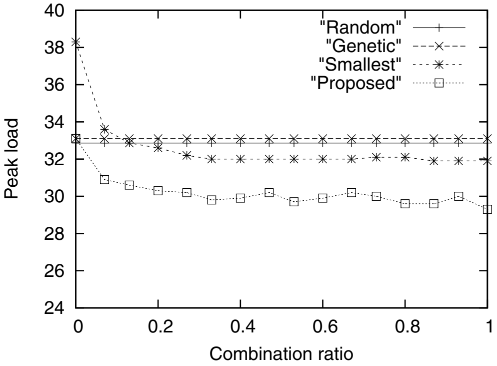


Fig. 6. Heuristic solution ratio

5 Conclusions

This paper has designed an efficient EV charging scheduler which combines the time-efficiency of heuristic-based approaches and the evolutionary iteration of genetic algorithms. It selects the initial population of a genetic algorithm from a time-efficient heuristic and then regular genetic operations are applied. The scheduler is aiming at reducing the peak power consumption in a charging station while meeting the time constraint of all charging tasks. By this scheduler, the concentrated charging problem can be relieved for the large deployment of EVs while a charging station can provide a reservation service to EVs, which can make a routing plan according to the interaction with charging stations. The performance measurement result obtained from a prototype implementation shows that our scheme outperforms *Random*, *Genetic*, *Smallest* schemes, reducing the peak load for the given charging task sets by up to 4.9 %. As future work, we are going to design an EV telematics framework which includes an efficient path planning scheme capable of integrating a charging schedule.

References

1. Gellings, C.: The Smart Grid: Enabling Energy Efficiency and Demand Response. The Fairmont Press (2009)
2. Lopes, J., Soares, F., Almeida, P.: Integration of Electric Vehicles in the Electric Power System. Proceedings of the IEEE, 168–183 (2011)
3. Markel, T., Simpson, A.: Plug-in Hybrid Electric Vehicle Energy Storage System Design. In: Advanced Automotive Battery Conference (2006)
4. Shao, S., Zhang, T., Pipattanasomporn, M., Rahman, S.: Impact of TOU Rates on Distribution Load Shapes in a Smart Grid with PHEV Penetration. In: Transmission and Distribution Conference and Exposition, pp. 1-6 (2010)
5. Facchinetti, T., Bibi, E., Bertogna, M.: Reducing the Peak Power through Real-Time Scheduling Techniques in Cyber-Physical Energy Systems. In: First International Workshop on Energy Aware Design and Analysis of Cyber Physical Systems (2010)
6. Togan, V., Dalgoglu, A.: An Improved Genetic Algorithm with Initial Population Strategy and Self-Adaptive Member Grouping. Computer & Structures, 1204–1218 (2008)
7. Lee, J., Kim, H., Park, G., Jeon, H.: Fast Scheduling Policy for Electric Vehicle Charging Stations in Smart Transportation. In: ACM Research in Applied Computation Symposium, pp. 110–112 (2011)
8. Morrow, K., Karner, D., Francfort, J.: Plug-in Hybrid Electric Vehicle Charging Infrastructure Review. In: Battelle Energy Alliance (2008)
9. Diaz-Gomez, P., Hougan, D.: Initial Population for Genetic Algorithms: A Metric Approach. In: International Conference on Genetic and Evolutionary Methods, pp. 43–49 (2007)
10. Toepfer, C.: SAE Electric Vehicle Conductive Charge Coupler, SAE J1772. Society of Automotive Engineers (2009)
11. Sortomme, E., Hindi, M., MacPherson, S., Venkata, S.: Coordinated Charging of Plug-in Hybrid Electric Vehicles to Minimize Distribution System Losses. IEEE Transactions on Smart Grid, 198–205 (2011)
12. Lee, J., Park, G.-L., Kang, M.-J., Kwak, H.-Y., Lee, S.J.: Design of a Power Scheduler Based on the Heuristic for Preemptive Appliances. In: Nguyen, N.T., Kim, C.-G., Janiak, A. (eds.) ACIIDS 2011, Part I. LNCS (LNAI), vol. 6591, pp. 396–405. Springer, Heidelberg (2011)
13. Lee, J., Park, G.-L., Kwak, H.-Y., Jeon, H.: Design of an Energy Consumption Scheduler Based on Genetic Algorithms in the Smart Grid. In: Jędrzejowicz, P., Nguyen, N.T., Hoang, K. (eds.) ICCCI 2011, Part I. LNCS, vol. 6922, pp. 438–447. Springer, Heidelberg (2011)

Self-Organizing Reinforcement Learning Model

Chang-Hsian Uang, Jiun-Wei Liou, and Cheng-Yuan Liou

Department of Computer Science and Information Engineering
National Taiwan University
cyliou@csie.ntu.edu.tw

Abstract. A motor control model based on reinforcement learning (RL) is proposed here. The model is inspired by organizational principles of the cerebral cortex, specifically on cortical maps and functional hierarchy in sensory and motor areas of the brain. Self-Organizing Maps (SOM) have proven to be useful in modeling cortical topological maps. The SOM maps the input space in response to the real-valued state information, and a second SOM is used to represent the action space. We use the Q-learning algorithm with a neighborhood update function, and an SOM for Q-function to avoid representing very large number of states or continuous action space in a large tabular form. The final model can map a continuous input space to a continuous action space.

Keywords: Reinforcement learning, Self-Organizing Maps, Q-learning, Unsupervised learning.

1 Introduction

Machine Learning can be generally classified as *supervised*, *unsupervised* and *reinforcement learning* (RL). Supervised learning requires clear input and output form for comparison, and the goal is to construct a mapping from one to the other. Unsupervised learning has no concept of mapping, and only process input data to find out the potential classification. In contrast, RL uses a scalar reward signal to evaluate input-output pairs and through trial and error to optimize the selected action for each input. RL can be considered as an intermediary between supervised and unsupervised learning. This learning method is most suitable for resolving an optimal input-output mapping without prior knowledge. These three learning paradigms are often combined to solve problem such as auto-associative multi-layer perceptron [1] and SRV units [2].

Q-learning is a reinforcement learning technique relied on learning an action-value function that gives the expected utility of taking a given action in a given state and following by a fixed policy thereafter. One of the strengths of Q-learning is ability of comparing the expected utility of the available actions without requiring a model of the environment. Q-learning is an iterative, incremental and interactive algorithm with simple update rule for Q-value, easy implementation and clear representation. Computational features in Q-learning seem to make neural networks a natural implementation choice for RL applications [3]. To combine RL theory with the

J.-S. Pan, S.-M. Chen, N.T. Nguyen (Eds.): ACIIDS 2012, Part I, LNAI 7196, pp. 218–227, 2012.

generalization offered by connectionism has been shown to provide very encouraging practical applications [4]. This article specifically addresses on the problem of representation and generalization of continuous action spaces in RL problem.

2 Reinforcement Learning

RL theory has provided an exhaustive framework for solving Markov decision problems. The target of problem is to maximize a scalar reward signal on the state transitions when the fundamental model dynamics are unknown. These methods are classified as Temporal Difference (TD) learning, and Q-learning is one of the most powerful methods.

As explained by the two pioneers of reinforcement learning, Richard S. Sutton and Andrew G. Barto, in their Reinforcement Learning book (1998): *Reinforcement Learning is best understood by stating the problem that we want to solve* [5]. The problem is learning to achieve a goal solely from interaction with the environment. The decision maker or learning element of RL is called an agent. The interactions between agent and environment are depicted in

Fig. 1. The agent selects actions and the environment reacts, changes the state and provides rewards to the agent indicating the degree of goodness of the action made by the agent. The mission of agent is to maximize the rewards received with training and errors.

To evaluate the goodness of a selected action within a definite state, a value function $(s, a) \rightarrow V(s, a)$ is defined which maps state-action pairs to a numerical value. While interacting with the environment, the agent must update this function to reflect how well the actions chosen progress based on the rewards received in the environment. How to update the values of the state-action pairs is pivotal issue of the RL [6].

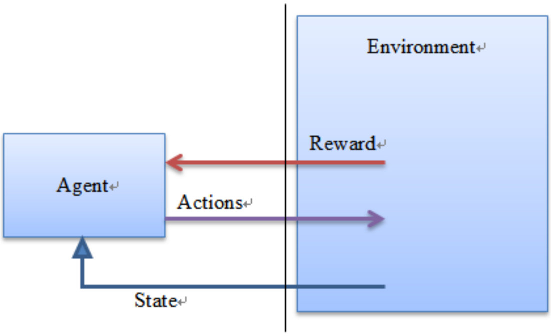


Fig. 1. Agent-Environment Interaction: the agent selects actions at each time step. The environment updates the state and defines a numerical reward.

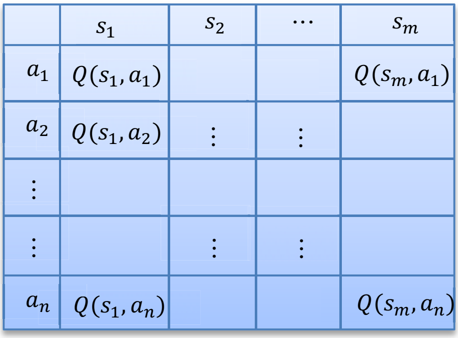
2.1 Q-Learning

Q-learning behaves as follows: an agent tries an action at a particular state, and then evaluates its consequences in terms of the received reward immediately after taking an action [7].

Q-learning requires a value table as in Fig. 2, where each entry represents the maximum expected reward from the corresponding state-action pair. The entries of the table are called as Q-values, and optimization is then a simple matter of selecting the action at each time-step with the highest Q-value for the current state. The only requirement for convergence comes from constantly updating all state-action pairs. Q-learning can be divided into following steps:

1. observe current state s_t
2. select and perform action a_t
3. observe the next state s_{t+1}
4. receive an immediate reward r_t
5. update the $Q(s_t, a_t)$ values using a learning factor according to the rule:

$$Q(s_t, a_t) \leftarrow Q(s_t, a_t) + \alpha \left[r_t + \gamma \max_a Q(s_{t+1}, a) - Q(s_t, a_t) \right] \tag{1}$$



	s_1	s_2	$\cdots$	s_m
a_1	$Q(s_1, a_1)$			$Q(s_m, a_1)$
a_2	$Q(s_1, a_2)$	$\vdots$	$\vdots$	
$\vdots$				
$\vdots$		$\vdots$	$\vdots$	
a_n	$Q(s_1, a_n)$			$Q(s_m, a_n)$

Fig. 2. Q-table: each table entry contains the predicted-Q value for each state-action pair

3 The Self-Organising Map

The Self-Organising Map (SOM) was first conceived by Kohonen in 1982. Kohonen describes the SOM as an "ordered, nonlinear, smooth mapping of high-dimensional input data manifolds onto the elements of a regular, low-dimensional array" [8]. It has recursive nature, and during each recursive step, only a subset of models is re-organized. In Fig. 3, we present a schematic for a 2-dimensional SOM.

A weight vector $w^t = [w_1^t, w_2^t, \ldots, w_D^t]$ is associated with unit t, where D is the dimensionality of the input data. The target is to find a suitable set of weights for each unit so that the network models the distribution of the input data in the input space. In many cases, the input data has lower dimensional representation then the input data dimensionality. In these cases, the SOM can be used as a dimension reduction technique.

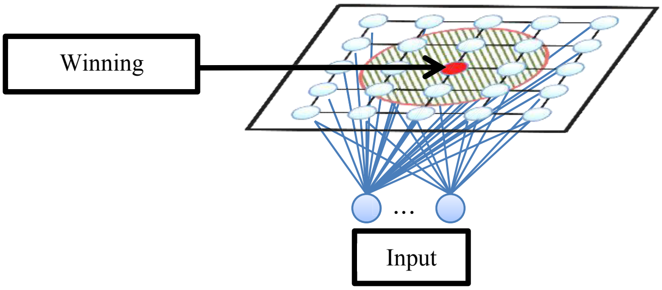


Fig. 3. 2-dimensional array of neurons conform to an SOM. The shaded circle denotes the neighborhood of the winning neuron.

The unit with the shortest Euclidean distance from input vector x is considered as the winner because of characterizing the current input most closely. The weights of the winner unit are immediately updated towards the input:

$$w^{winner} = w^{winner} + \beta(x - w^{winner}), \tag{2}$$

where β is the learning rate. Neighbors of the winner unit are also processed using similar formula except the modification term is further multiplied by a decay parameter according to the distance of those neighbors from the winner.

The weights of the map are initialized to random values. The above operation is iterated for each input vector in the dataset, and effectively results in a competition between different regions of the input space for units on the map. Compact regions of the input space will attract more units than sparse ones. Neighborhood learning also promotes topology preservation such that units closed to each other in the topology of the input space will end up to be closed to each other in the weight space.

4 The Model

The problem can be considered as searching for an optimal mapping from a continuous input space to a continuous action space. An optimal mapping means maximizing the expected total reward within each possible state. The model uses an SOM to quantize a continuous input space into a discrete representation. The SOM maps input data in response to the real-valued state information, and the index of each unit is interpreted as a discrete state in the Q-table. A second SOM is the same as above and used to represent the action space, with each unit of this second map corresponding to a discrete action in the Q-table.

The first SOM is called *input map,* inhabits in the state space and attempts to represent the input space at the highest resolution in the most active regions. The second SOM is called *action map*, which inhabits in the action space. The second SOM must be explored by trial and error to discover the highest reward for the whole range of observed inputs. The following algorithm is used to achieve this exploration: for any real-valued state vector, the unit of the input map with smallest Euclidean distance from that state vector is identified as the winner; next, one of the units of the action map is selected according to the Q-learning criterion—i.e. the one with the highest Q-value for the current state if *exploiting*, and a random action if *exploring*. This winning action unit is used as the basis for the real-valued action. We can get the action weight vector as the *proposed action*. The proposed action is then perturbed by random noise and becomes *perturbed action* for the actual output. If the reward received with the perturbed action is better than the estimated expected return associated with the winning state–action pair, then the exploration in the action map appears to be successful, so the action map is updated towards the perturbed action. Otherwise, no learning takes place in the action map. In any case, the Q-value of the winning state–action pair is updated towards the actual one-step corrected return.

The algorithm can be interpreted as standard Q-learning with the discrete states and the discrete actions. Besides the topology preserving nature of the SOM, a simple amendment to expedite the algorithm is not only to update the Q-value of the winning state-action pair, but to update *every* state–action pair towards this value proportional to the product of the two neighborhood functions (of the input and action maps). We call this *neighborhood Q-learning* [9]. The complete algorithm is summarized as below for the case of a 2-dimensional input space and a 2-dimensional action space:

1. Present an input vector, I, to the system, and identify the winner in the *input map*, s_j.
2. Identify a unit in the *action map*,

   $$a_k =. \begin{cases} \text{One with best Q-value for state, } s_j & \text{with probability } 1-p \\ \text{Random action} & \text{with probability } p \end{cases}$$

3. Identify the *proposed action* as the weights of unit a_k, $\langle u_{k1}, u_{k2} \rangle$.
4. Independently perturb each element of the proposed action by a small random noise to yield a new *perturbed action*:

   $$\langle u'_{k1}, u'_{k2} \rangle = \langle u_{k1} + random(-1,1) \times \varepsilon, u_{k2} + random(-1,1) \times \varepsilon \rangle$$

5. Output the perturbed action vector $\langle u'_{k1}, u'_{k2} \rangle$.
6. Receive a reward r from the environment.

7. If $r + \gamma max_i Q(s_j', a_i) > Q(s_j, a_k)$, then the perturbed action appears to be an improvement over the existed action, so update the action map towards the perturbed action, $\langle u_{k1}', u_{k2}' \rangle$, according to the usual SOM update rule:

$$u_{m1} := u_{m1} + \lambda_A \times \psi_A(k, m, N_A)(u_{k1}' - u_{m1})$$
$$u_{m2} := u_{m2} + \lambda_A \times \psi_A(k, m, N_A)(u_{k2}' - u_{m2})$$

   for all action units m.

8. Update *all* Q-values towards the corrected return proportionally to the Q-learning rate and the product of the two neighborhoods (neighborhood Q-learning):

$$Q(s_m, a_n) := Q(s_m, a_n) + \alpha \times \psi_S(j, m, N_S) \times \psi_A(k, n, N_A) \times (r + \gamma\, max_i\, Q(s_j', a_i) - Q(s_m, a_n))$$

   for all state units m, and actions units n.

9. Update the input map towards vector I according to the usual SOM update rule:

$$w_{m1} := w_{m1} + \lambda_S \times \psi_S(j, m, N_S)(I_1 - w_{m1})$$
$$w_{m2} := w_{m2} + \lambda_S \times \psi_S(j, m, N_S)(I_2 - w_{m2})$$

   for all state units m.

10. Return to 1.

where s_j' is the state immediately after s_j, λ_A is the learning rate of the action map, λ_S is the learning rate of the input map, α is the Q-learning rate, u_{ki} is the ith weight of the kth unit of the action map (w are the weights of the input map), $random(-1,1)$ yields a random number selected uniformly from the range $[-1,1]$, ε controls the amount of exploration in the action space, and $\psi_S(j, m, N_S)$ is the value of the neighbourhood function of the input map at unit m given winning unit j and a neighborhood size N_S (similar for ψ_A only for the action map). A simple linear neighborhood is used so that $\psi_S(j, m, N_S) = \max(0, 1 - (d/(N_S + 1)))$ where d is the distance between units j and m in the topology of the map. All these parameters must be set empirically and annealed together throughout learning. The model is illustrated in Fig. 4.

5 Experiments

In this experiment, the control problem requires a mapping learned from a continuous 2-dimensional state space to a continuous 2-dimensional action space. In Fig. 5, a goal is generated at random on the circle shown, and the coordinates of the goal are provided as input to the learning agent. The agent must output an adequate set of joint angles so that the tip of the arm touches the goal. After an action is taken, reward is immediate and is simply the negative of the distance between the tip of the arm and the goal. As shown in Fig. 6, the main task of agent is to learn a mapping from goal space to arm space. Note that this is not supervised learning because the reward does not contain direction information to the agent.

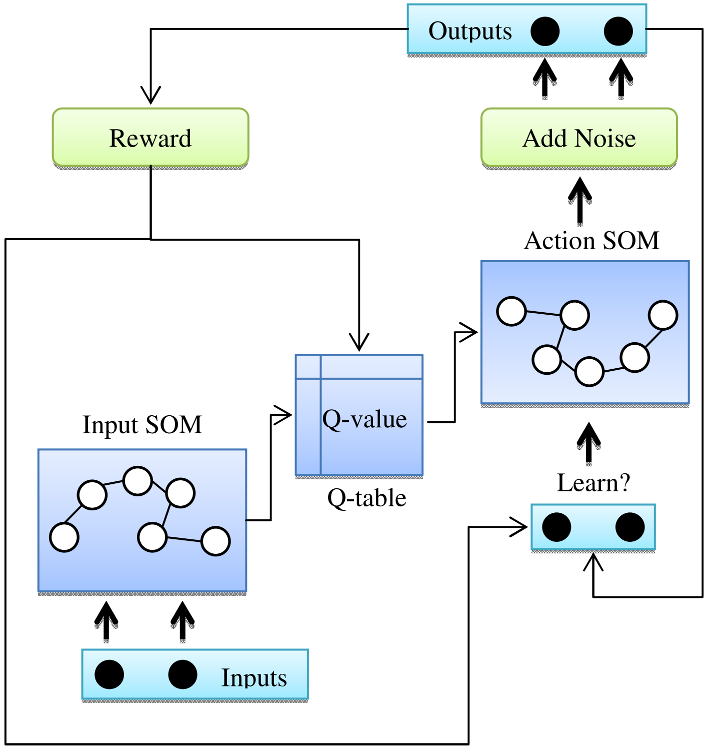


Fig. 4. The proposed learning model applied Q-learning and SOM for continuous inputs and outputs

Table 1 shows the set of empirical parameters used to achieve the results in Fig. 7. Performance depends crucially on the annealing speed for exploration and plasticity. There are trade-offs between fast convergence of exploratory noise and sufficient exploration for adequate search of the action space. The graph shown is for the fastest annealing schedule which can significantly affect the final mean performance. Fig. 8 shows the input and action maps after learning on a typical trial. The input map has responded to the input distribution of goal positions, and the action map, through a process of trial and error, has also learned to represent those actions which offer optimal solutions to the states represented in the input map. The strategy which mapping input units to action units is indicated by the shading. It is obvious that topology is preserved not only in the two maps, but also in the Q-table. The performance graph from Fig. suggests that the action units indeed occupy higher rewarded regions of the action space for the whole range of inputs.

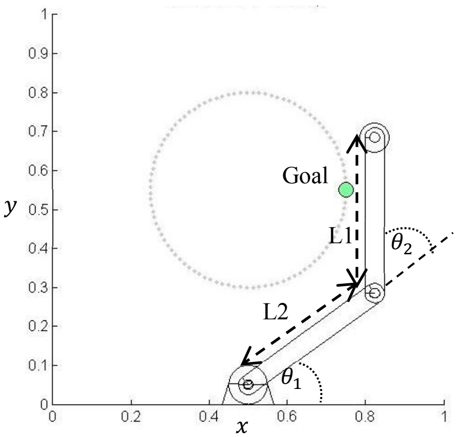


Fig. 5. A simulation with two-joint arm residing inside the unit square. The base of the arm is fixed at (0.5, 0), and the position of the end point of the arm is defined by the lengths of the two arm segments (*L1* and *L2*), and two relative angles, θ_1 and θ_2, are measured in radians and restricted to the range $[-\pi,\pi]$. Note that θ_1 is given relative to the 'x-axis'.

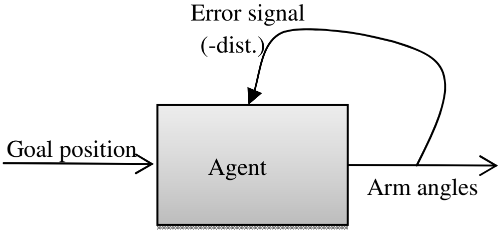


Fig. 6. The task is to learn a mapping from goal space to joint-angle space using an immediate reward signal. The outputs of the agent are θ_1 and θ_2. Note that these are not angles to move through, but angles to move to. θ_1 and θ_2 uniquely describe the arm configuration.

Table 1. Empirical parameters for the proposed model's application to the multi-joint arm problem. Initially, all Q-values are set to zero and all action SOM weights are generated uniformly within the range $[-\boldsymbol{\pi},\boldsymbol{\pi}]$, and input SOM weights in the range [0,1]. $\boldsymbol{t}$ is the time-step in which a single cycle of the algorithm is performed.

Parameter	Value
Input map size	50 × 1 units
Action map size	50 × 1 units
Input map neighborhood size, N_S	$10\ \times f(t)$
Action map neighborhood size, N_A	$10\ \times f(t)$
Q-learning rate, α	$f(t)$
Discount factor, γ	0
Learning rate of input map, λ_S	$f(t)$

Table 1. (*continued*)

Learning rate of action map, λ_A	$f(t)$
Probability of Q-learning exploration, p	$f(t)$
Max. exploration distance around action unit, ε	$f(t)$
Annealing schedule, $f(t)$	$1/\left(\frac{t}{1000}+1\right)$

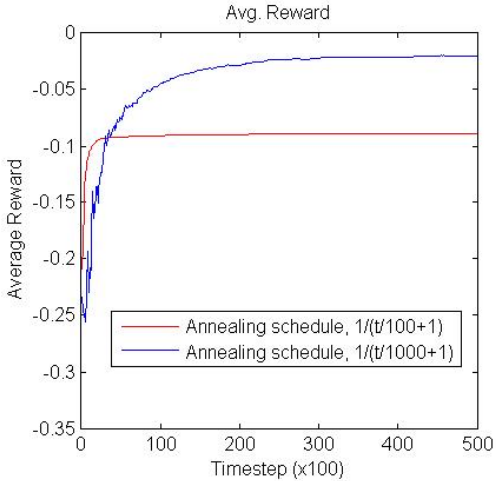


Fig. 7. Average reward against time using the parameters of Table 1. Each data point is averaged over the last 1000 time-steps for each trial, and the plot shows the mean of these data points over 20 independent trials.

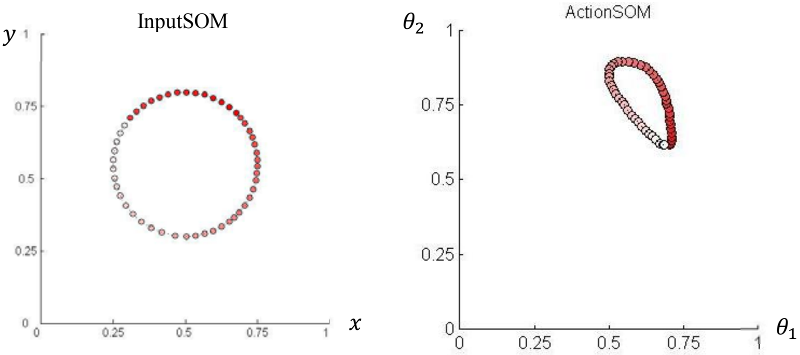


Fig. 8. Input map plotted in input space after learning on a typical trial. Units are shaded according to the action unit with the highest Q-value for that input unit. Action map plotted in action space after learning on a typical trial. Each angle is normalized to the range [0,1] (from the range $[-\pi, \pi]$). Units are shaded according to their topological index in the action map.

6 Conclusion

A model has been presented for representation and generalization in model-less RL. The proposed model is based on the SOM and one-step Q-learning which provides several desirable properties: real-valued states and actions are adapted dynamically; a real-valued and potentially delayed reward signal is accommodated. The core of the model is RL theory which is adhered to an explicit notion of estimated expected return. The model has both an efficient learning and action selection phase, and as long as the nature of the desired state–action mapping is unknown beforehand. The topology preserving property of the SOM has also been shown to be able to add useful constraints under constrained tasks, and to expedite learning through the application of neighborhood Q-learning.

The drawbacks of this model are the theoretical possibility that arbitrarily poor actions could be observed by the system given a sufficiently malicious reward function, and the inherent scalability issues resulting from the fact that the representation uses local parameters. Distributed models may offer a solution to the scalability issue, but these models have been shown to introduce their own problems pertaining to flexibility, stability, and robustness in the face of unpredictably distributed training data. Smith (2001, Chapter 9) [10] suggest to combine local and distributed models may be of future interest to the kinds of applications considered here.

Acknowledgments. This work was supported by National Science Council under project number NSC 100-2221-E-002-234-MY3.

References

1. Rumelhart, D.E., Hinton, G.E., Williams, R.J.: Learning internal representations by error propagation. In: Parallel Distributed Processing, vol. 1. MIT Press, Cambridge (1986)
2. Gullapalli, V.: A stochastic reinforcement learning algorithm for learning real-valued functions. Neural Networks 3, 671–692 (1990)
3. Smith, J.A.: Applications of the self-organizing map to reinforcement learning. Neural Networks 15, 8–9 (2002)
4. Tesauro, G.J.: Practical issues in temporal difference learning. Machine Learning 8, 257–277 (1992)
5. Sutton, R.S., Andrew, G.B.: Reinforcement Learning. MIT Press (1998)
6. Luis, R.S.: The Hierarchical Map Forming Model. Master's thesis, Department of Computer Science and Information Engineering, College of Electrical Engineering and Computer Science, National Taiwan University (2006)
7. Watkins, C.J., Dayan, P.: Technical Note: Q-Learning. Machine Learning 8, 22 (1992)
8. Kohonen, T.: Self organization and associative memory, 2nd edn. Springer, Berlin (1987)
9. Smith, J.A.: Applications of the self-organizing map to reinforcement learning. Neural Networks 15, 8 (2002)
10. Smith, A.J.: Dynamic generalization of continuous action spacesin reinforcement learning: A neutrally inspired approach. PhD dissertation, Division of Informatics, Edinburgh University, UK (2001)

Facial Feature Extraction and Applications: A Review

Ying-Ming Wu[1], Hsueh-Wu Wang[2], Yen-Ling Lu[3],
Shin Yen[4], and Ying-Tung Hsiao[2]

[1] Department of Electrical Engineering,
Tamkang University, 25137 Taipei, Taiwan
[2] Department of Digital Technology Design,
National Taipei University of Education, 10671 Taipei, Taiwan
[3] Central Personnel Administration, Executive Yuan, 10051 Taipei, Taiwan
[4] Department of Computer Science, National Taipei University of Education,
10671 Taipei, Taiwan
{hwwang,ythsiao}@tea.ntue.edu.tw
wuym@pchome.com.tw
anny@cpa.gov.tw
cindy30114@yahoo.com.tw

Abstract. Facial feature extraction plays an important step in automated visual interpretation and human face recognition. Detecting facial feature is a crucial role in a wide variety of application such as human computer interface, facial animation and face recognition, etc. The major objective of this paper is to review the recent developments on the methods of facial feature extraction. This study summaries different method for feature point extraction and their applications on face image identification and highlight the performance regarding these methods. The major goal of the paper is to provide a summary reference source for the researchers involved in facial feature extraction.

Keywords: Eye detection, face detection, face recognition, image processing, Gabor Wavelet Transform.

1 Introduction

Detection of facial features is a major concern in a wide variety of applications such as human computer interaction, facial animation, facial expression, face recognition, and face image database management. Detecting facial features is a pivotal step in these applications. Facial feature extraction is nothing but identifying exact location of different features on face which include detection of eyes, brows, mouth, nose, chin etc. Human being is able to recognize almost a thousand of faces in one's whole life and distinguish faces without any difficulty. Nevertheless, facial features extraction is difficult representation in computer programs because it requires differentiation among human faces which vary subtly from each. Moreover, building an identifying index enabling all features to be physically applied to the problem is not a simple task.

J.-S. Pan, S.-M. Chen, N.T. Nguyen (Eds.): ACIIDS 2012, Part I, LNAI 7196, pp. 228–238, 2012.

Analyzing and identifying facial features events are laborious when the methods adopted are primarily based on visual inspection. Accordingly, a method that can automatic classify facial features in an image must be developed.

Several novel and particularly successful object and object category detection and recognition methods based on image features, local descriptions of object appearance, have recently been proposed. Most of the research and development activities in face detection and identification known as the biometric authentication. In the recent researches to these topics the focus of interest has changed from the image recognition techniques toward the feature distinguish.

Many researches developed various methods for extraction and recognition of facial features on gray and color images. The studies on facial feature extraction continue to develop high accuracy, reduced complexity, high efficiency and less computational time approach.

2 Performance-Driven Facial Animation

Performance-driven facial animation (PDFA) is based on the performer for the facial expressions. PDFA use sensing devices in a virtual character to reconstruct facial expressions. Williams [1] proposed a prototype, can be divided into two main steps: feature tracking and 3D model construction. This method capture the performer's facial feature points and mapped into the corresponding points of the face. And then, after each movement for the performers do a track, and change the corresponding model. Therefore, the method based on the cyclic process is also known as feature retargeting. In the process of capturing the action, it is need to apply noise filter between frames, such as band-pass filter.

The prototype PDFA is slightly restricted during the process of capturing features in some periods of performance in the film. It cannot capture motion vectors form the role for mapping to the new virtual character directly to replicate the same action. It must be built a new one for duplication. In addition the movement between the frame coherence is lack of intelligent systems for analysis these changes. It is necessary to manually modify and fill for continuity. Also, the mapping of corresponding points cannot directly use the context before and after the action.

Khanam and Mufti [2] utilized the fuzzy system and MPEG-4 feature points to estimate the location of and employed open source software Xface to produce facial animation. The fuzzy system design a membership function based the gap value of each feature point locating in a different frame, and use fuzzy set of feature points to design fuzzy rules for determining the respective expression. Finally, this approach is according to the corresponding results to the animation, and enhances coherence between frames to making more natural animation, the process shown in Fig. 1, the results shown in Fig. 2.

Cosker et al. proposed the combination of expressions from multiple samples to generate a new expression image, such as selecting the mouth expression from face image A and the forehead expression of face image B to generate the new expression in C and then applying to the objective model [3]. Fragments of information about these features can be stored in a database for future use, diagram shown in Fig. 3. This method applied principal component analysis to create feature vectors, and use active appearance model as a fragment of the contour features to meet the shapes with different characteristics. Figure 4 shows the images on selection of characteristics by active appearance model.

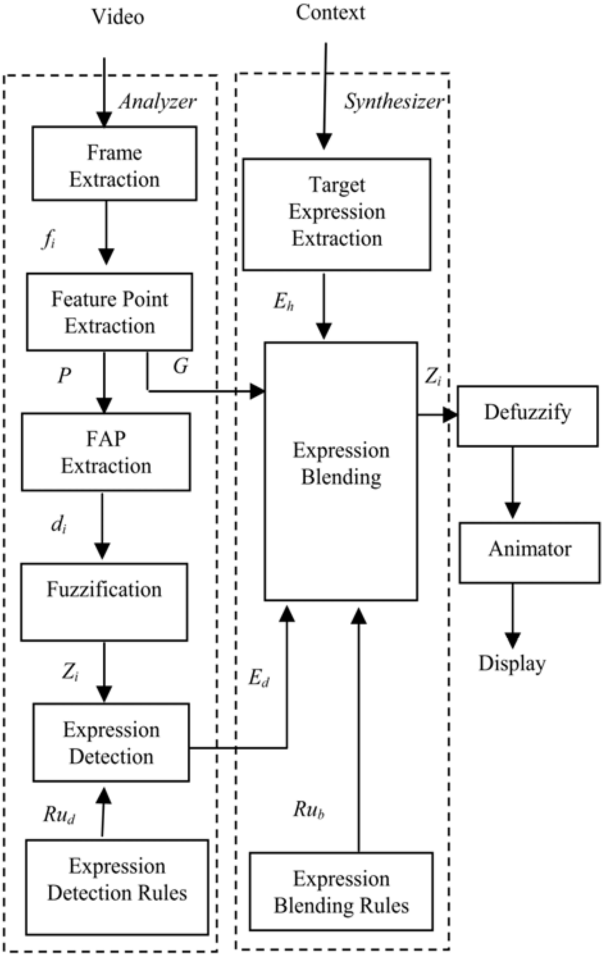


Fig. 1. The flowchart of fuzzy system for PDFA [2]

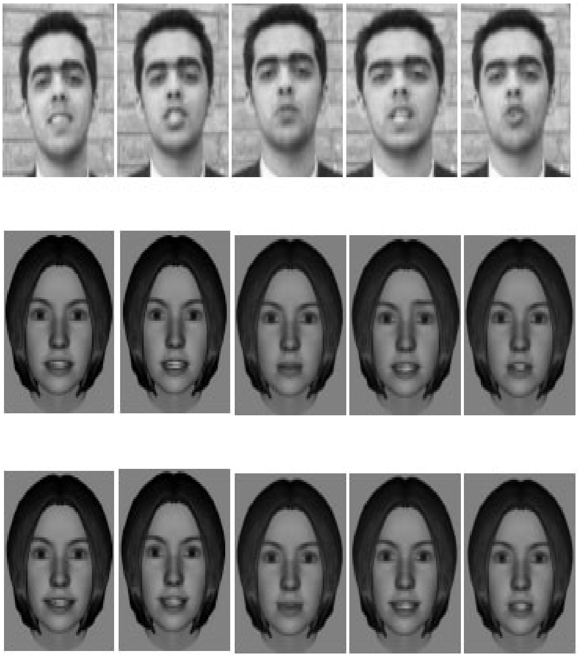


Fig. 2. The results of estimating action by using the fuzzy system [2]

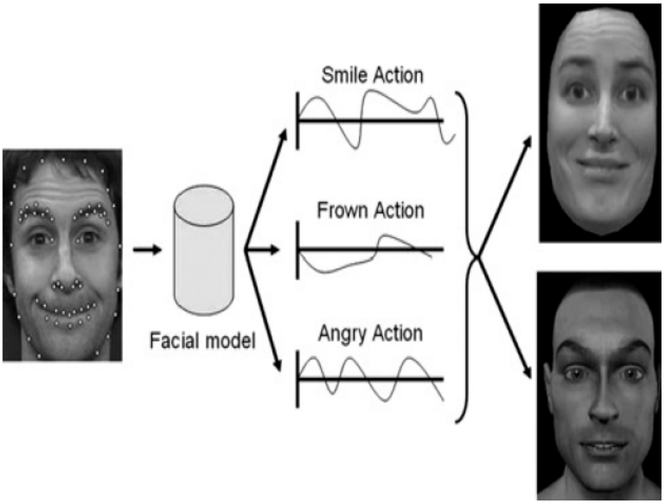


Fig. 3. Multiple expression image generating system [3]

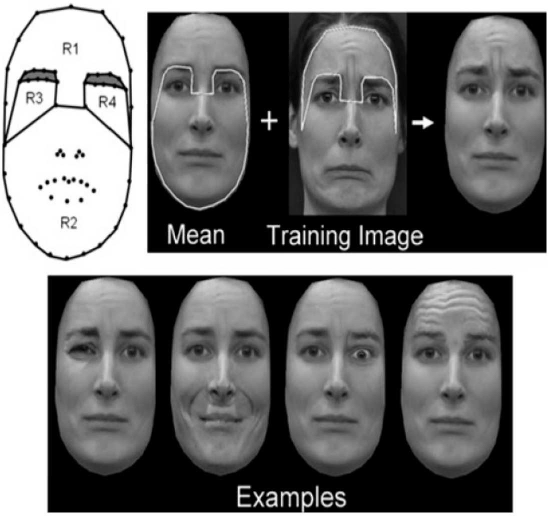


Fig. 4. The images on selection of characteristics by active appearance model [3]

3 Color Segmentation

Color segmentation is based on the unique color and background color information to separate images. Jiang, Yao and Jiang proposed the color detection method using skin probability map to set the probability threshold to cut the color and non-color area with very high true acceptance rate, but the false acceptance rate is also high[4]. Therefore, in order to reduce the false acceptance rate, color segmentation is joined the texture filter based on the Gabor wavelet transform-based to acquisition the texture feature. Although the use of texture filter reduces the false acceptance rate while it also reduces the rate of identification. In order to improve the rate of identification, using morphological markers in the control of the marker-controlled watershed segmentation analyze the previous texture image to calculate the mean deviation and standard deviation, and a preset threshold. And more, if the mean deviation and standard deviation smaller than the threshold value, the skin region can be determined. The process is roughly divided into three phases: SPM color filter, texture filter and mark-controlled watershed segmentation, as display in Fig. 5.

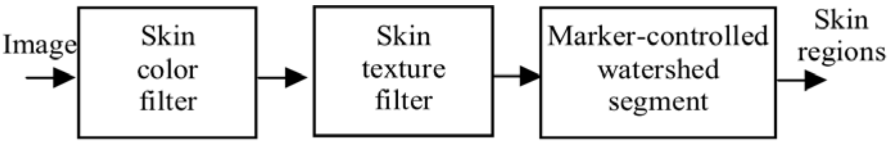


Fig. 5. The flow chart of color segmentation [4]

Phung, Bouzerdoum and Chai [5] utilized four classifiers to assess the accuracy of color segmentation based on four common image processing color spaces, i.e. RGB, HSV, YCbCr, CIE-Lab. 1) Piecewise linear decision boundary classifiers observe the distribution of image pixel color values and set a fixed color range. In this category of classifiers, skin and non-skin colors are separated using a piecewise linear decision boundary. 2) Bayesian classifier with the histogram technique applied the decision rule to consider a color pixel as a skin pixel. The decision rule is based on the a priori probabilities of skin and non-skin and various classification costs and is associated with a threshold determined empirically. The class-conditional pdfs can be estimated using histogram or parametric density estimation techniques. 3) Gaussian Classifiers. The class-conditional pdf of skin colors is approximated by a parametric functional form, which is usually chosen to be a unimodal Gaussian [6,7] or a mixture of Gaussians [8,9]. In the Gaussian Classifiers, A color pixel x is considered as a skin pixel when a threshold is smaller than the squared Mahalanobis distance. 4) The multilayer perceptron (MLP) is a feed-forward neural network that has been used extensively in classification and regression. A comprehensive introduction to the MLP can be found in [10]. Compared to the piecewise linear or the unimodal Gaussian classifiers, the MLP is capable of producing more complex decision boundaries.

The classification rates (CR) of the tested classifiers are shown in Table 1. The Bayesian and MLP classifiers were found to have very similar performance. The Bayesian classifier had a maximum CR of 89.79 percent, whereas the MLP classifier had a maximum CR of 89.49 percent. Both classifiers performed consistently better than the Gaussian classifiers and the piecewise linear classifiers. The classification rates (CRs) at selected points on the ROC curves are shown in Table 2. We observe that, at the histogram size of 256 bins per channel, the classification performance was almost the same for the four color spaces tested, RGB, HSV, YCbCr, and CIE-Lab. As our results show, such an expansion leads to more false detection of skin colors and reduces the effectiveness of skin segmentation as an attention-focus step in object detection tasks. Using lighting compensation techniques such as the one proposed by Hsu et al. [11] is probably a better approach to coping with extreme or biased lightings in skin detection.

Table 1. Classification Rates (CRs) of Skin Color Pixel Classifiers [5]

Classifier ID	CbCr-fixed	HS-fixed	GT plane-set	Baye-sian	2DG-pos	3DG-pos	3DG-pos/neg	3D-GM	MLP
FDR=10%	FDR= 29.09 CR= 75.64	FDR= 19.48 CR= 78.38	FDR= 18.77 CR= 82.00	88.75	82.37	85.27	88.01	85.23	88.46
FDR=15%				86.17	81.07	83.45	85.57	83.63	85.97
FDR=20%				82.97	78.85	80.84	82.47	81.29	82.84
CR_{max}				89.79	82.67	85.57	88.92	85.76	89.49
99% conf. Int. of CR_{max}	[75.58 75.70	[78.32 78.44]	[81.94 82.06]	[89.74 89.84]	[82.61 82.73]	[85.52 85.62]	[88.87 88.97]	[85.71 85.81]	[89.44 89.54]

Table 2. Classification Rates (CRs) of Eight Color Representations (Histogram Size = 256 Bins per Channel) [5]

	All channels				Only Chrominance Channels			
Color Representation	RGB	HSV	YCbCr	CIE-Lab	rg	HS	CbCr	ab
FDR = 10%	88.75	88.76	88.46	88.47	84.93	85.46	86.56	84.51
FDR = 15%	86.17	86.19	85.97	85.97	83.92	84.14	84.77	83.45
FDR = 20%	82.97	82.98	82.85	82.83	81.65	81.91	82.11	81.33
CR_{max}	89.79	89.80	89.41	89.43	84.95	85.58	86.75	84.52
99% conf. Int. of CR_{max}	[89.74 89.84]	[89.75 89.85]	[89.36 89.46]	[89.38 89.48]	[84.90 85.00]	[85.53 85.63]	[86.70 86.80]	[84.47 84.57]

4 Feature Detection

Feature detection employs a particular description to distinguish or retrieve the information of interest within images or blocks, such as edges, corners, colors, etc., for the establishment of the corresponding characteristic value (e.g. eigenvalue) of information in order to facilitate general search. Then the connective step is to adopt the appropriate image classifier to identify the existence of the same characteristics among images or the features of value in the approximate range of information. These steps are known as the feature extraction. Feature detection in the spatial domain and frequency domain is with different methods used facial features on the detection and capture in the relevant literature for the introduction in the following.

Spatial domain is the image coordinate space. The pixel value at each coordinate is the intensity of the point. The applications on the spatial domain in practice around this area are almost the computing of image pixels [12]. Song et al. [13] proposed to distinguish the expression on the face wrinkles caused by changes in image intensity based on skin deformation parameters as the face recognition features. Figure 6 shows the recognizing facial expression on the eight regions (patch) defined by the MPEG-4 feature points where the cross marks represent the feature points of the FAP. Lines connected to each area using image ratio features to calculate the eigenvalues of each region to estimate the skin deformation parameters. Image ratio features can affect the accuracy of detection on handling images due to changes in light characteristics. The affecting of light conditions has always been an important issue mostly on the feature detection applications [14].

Figure 7 is an example of using the image ratio features where (a) and (c) are the expression and nature face images, respectively and the blocks of the same region is characterized by (b). The identifications of the image can base on comparing different images with the image ratio features.

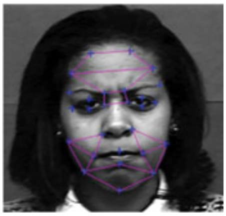


Fig. 6. The facial expression on the eight patch defined by MPEG-4[14]

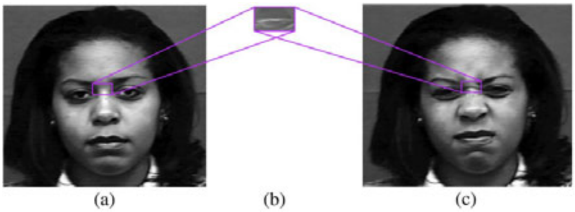


Fig. 7. The image ratio features [14]

Figure 8 shows the results of the recognition rate on the SDP, FAP and mixed approach. Figure 9 exhibits the recognition rate profiting from the using FAP and mixed features. The experimented results display the characteristics of the overall recognition rate of SDP with a noticeable improvement.

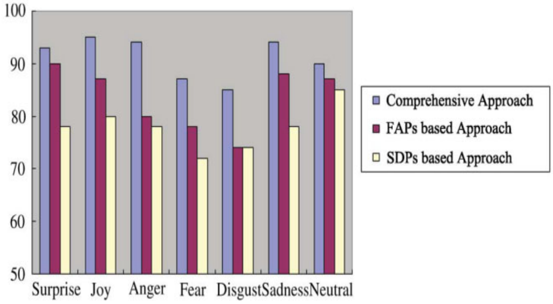


Fig. 8. Comparison of the recognition rate of the SDP, FAP and comprehensive approaches [13]

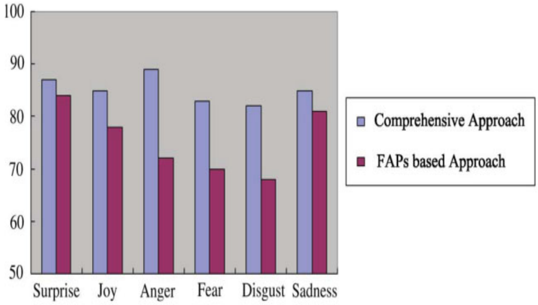


Fig. 9. The recognition rate profiting from the FAP and comprehensive approaches in contrast with SDA approach [13]

In contrast to the spatial domain, the image is treated as a period function in frequency domain. The image is represented by frequency and phase angle through Fourier transform. Ilonen et al. proposed the multi-resolution Gabor feature for various detection applications on facial images, such as the recognition on face, license plate to extract the characteristics of the objective images [15]. Gabor filter feature is captured by Gabor filter and designed as an image processing operator, called the simple Gabor feature. In order to be able to handle more complex problems, such as looking at the characteristics of low-contrast images, Gabor filter is integrated with multi-resolution space for forming a more efficient operator. Gabor filter translates the pixel information of the image into different frequency band. The low frequency of the image represents slowly changing information. In general, it is the gray image areas, such as the tone of similar background.

As to the high frequency band, the high frequency of the image illustrates the rapid changes in information. Usually, it is the area with high changes of the image, such as the details of information objects, edges or the noise of image [10,12]. For two-dimensional Gabor filter image processing is the Gaussian low-pass filter represented in the form of complex plane [10, 15, 16]. Figure 10 illustrates the extracting features and position on the face and license plate by using the Gabor filter. As to the optimal parameter adjustment and effectiveness in its mathematical proof can be seen in [15,16].

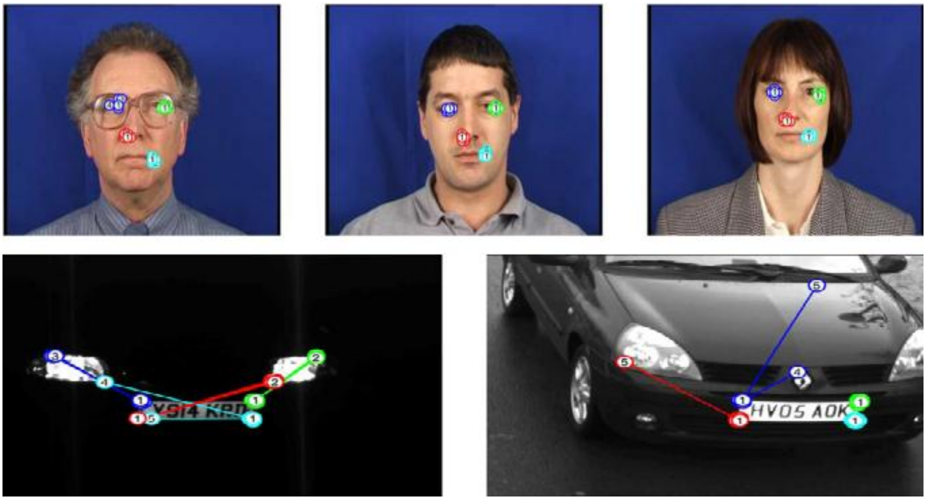


Fig. 10. The extracting features and position on the face and license plate by using the Gabor filter

5 Image Registration

Image Registration is from different viewpoints and at different time, and/or different sensors on the same picture according to their geometric relationship or the calibration model by overlapping the multiple images for Image fusion [7]. Image registration is usually applied to Image analysis such as comparing the characteristics of information to observe the characteristic changes on different images for estimating the possible change direction. Image registration is divided into base Image and reference Image (input Image and sensed image) for further process.

Image registration is widely used in remote sensing, medical imaging, computer vision etc. There is not a comprehensive method for image registration because the image source and target features are quite extensive and with various senses. In general, it is need to establish a specific mode of image registration for different geometric changes, noise filtering, and calibration accuracy. The majority of the registration methods consists of the following four steps, feature detection, feature matching, transform model estimation as well as image resampling and transformation to synthesis the new image by overlap the reference image and the sensed image.

Figure 11 presents the four steps of image registration: top row—feature detection (corners were used as the features in this case). Middle row—feature matching by invariant descriptors (the corresponding pairs are marked by numbers). Bottom left—transform model estimation exploiting the established correspondence. Bottom right—image resampling and transformation using appropriate interpolation technique.

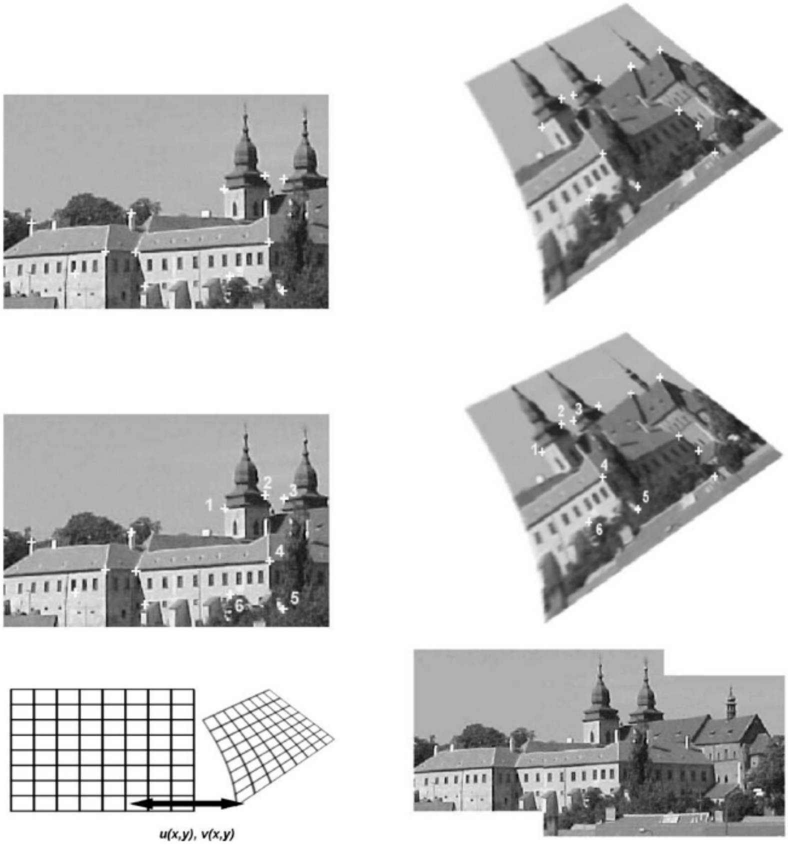


Fig. 11. The four steps of image registration [7]

6 Conclusions

This paper presents different methods for feature point extraction and highlights their performance. Various applications on feature point extraction are also summarized in this study to provide a guide reference source for the researchers involved in facial feature extraction and their applications.

References

1. Williams, L.: Performance-Driven Facial Animation. ACM SIGGRAPH Computer Graphics 24(4), 235–242 (1990)
2. Khanam, A., Mufti, M.: Intelligent Expression Blending for Performance Driven Facial Animation. IEEE Transactions on Consumer Electronics 53(2), 578–583 (2007)
3. Cosker, D., Borkett, R., Marshall, D., Rosin, P.L.: Towards Automatic Performance-Driven Animation between Multiple Types of Facial Model. IET Computer Vision 2(3), 129–141 (2008)
4. Jiang, Z., Yao, M., Jiang, W.: Skin Detection Using Color, Texture and Space Information. In: Proc. of the Fourth International Conf. on Fuzzy Systems and Knowledge Discovery (FSKD 2007), August 24-27, vol. 3, pp. 366–370 (2007)
5. Phung, S.L., Bouzerdoum, A., Chai, D.: Skin Segmentation Using Color Pixel Classification: Analysis and Comparison. IEEE Transactions on Pattern Analysis and Machine Intelligence 27(1), 148–154 (2005)
6. Yang, J., Waibel, A.: A Real-Time Face Tracker. In: Proc. IEEE Workshop Applications of Computer Vision, pp. 142–147 (December 1996)
7. Menser, B., Wien, M.: Segmentation and Tracking of Facial Regions in Color Image Sequences. In: SPIE Visual Comm. and Image Processing, vol. 4067, pp. 731–740 (June 2000)
8. Greenspan, H., Goldberger, J., Eshet, I.: Mixture Model for Face Color Modeling and Segmentation. Pattern Recognition Letters 22, 1525–1536 (2001)
9. Yang, M.-H., Ahuja, N.: Gaussian Mixture Model for Human Skin Color and Its Applications in Image and Video Databases. In: SPIE Storage and Retrieval for Image and Video Databases, vol. 3656, pp. 45–466 (January 1999)
10. Theodoridis, S., Koutroumbas, K.: Pattern Recognition, 4th edn. Academic Press, Burlington (2009)
11. Hsu, R.L., Abdel-Mottaleb, M., Jain, A.K.: Face Detection in Color Images. IEEE Transactions on Pattern Analysis and Machine Intelligence 24(5), 696–706 (2002)
12. Gonzalez, R.C., Woods, R.E.: Digital Image Processing, 2nd edn. Prentice Hall, New Jersey (2002)
13. Song, M., Tao, D., Liu, Z., Li, X., Zhou, M.: Image Ratio Features for Facial Expression Recognition Application. IEEE Transactions on Systems, Man, and Cybernetics – Part B: Cybernetics 40(3), 779–788 (2010)
14. Mitra, S., Acharya, T.: Gesture Recognition: A Survey. IEEE Transactions on Systems, Man, and Cybernetics – Part C: Applications and Reviews 37(3), 311–324 (2007)
15. Ilonen, J., Kamarainen, J.K., Paalanen, P., Hamouz, M., Kittler, J., Kälviäinen, H.: Image Feature Localization by Multiple Hypothesis Testing of Gabor Features. IEEE Transactions on Image Processing 17(3), 311–325 (2008)
16. Ilonen, J., Kamarainen, J.K., Kälviäinen, H.: Efficient Computation of Gabor Features, Research Rep. 100, Dept. Inf. Technol., Lappeenranta Univ. Technol., Finland (2005)

An Intelligent Infant Location System Based on RFID

Shou-Hsiung Cheng

Department of Information Management,
Chienkuo Technology University, Changhua 500, Taiwan
shcheng@ctu.edu.tw

Abstract. This study proposes a straightforward and efficient infant location system to reduce the potential risks of the theft and misuse hold. The system can recognize without difficulty different locations of newborn babies which are attached wristband active RFID tags. The system can accurately recognizes the locations of newborn babies by using decision tree classifiers after the active RFID readers has received different intensity of electromagnetic waves transmitted by active RFID tags.

Keywords: infant location system, RFID, decision tree.

1 Introduction

As the characteristics of newborn babies are similar and the lack of expression of ability, often cause of neonatal injury due to human negligence or error identified. Infancy is the most vulnerable stage, a variety of injuries can cause permanent sequelae or death. How to enhance the identification of newborns and to prevent misuse hold are the important projects in neonatal care. Infant hearthcare with RFID discarding the traditional paper job is becoming more and more common as hospitals in today's competitive environment. Wristband active RFID tags can be attached to infants shortly after birth.

In recent years, Radio Frequency Identification (RFID) technology has been widely accepted in hospitals to track and locate newborn babies. Hightower [1] was the first to use RFID technology to build the indoor positioning system. He uses RSS measurements and proposed SpotON system to estimate the distance between a target tag and at least three readers and then applies trilateration on the estimated distances. Lionel [2] proposed the LANDMARC method based on RFID to build indoor positioning system. LANDMARC system have a greater read range and response capability of the active sensor tags compared to SpotON. Wang et al. [3] propose a 3-D positioning scheme, namely passive scheme, which relies on a deployment of tags and readers with different power levels on the floor and the ceiling of an indoor space and uses the Simplex optimization algorithm for estimating the location of multiple tags. Stelzer et al. [4] uses reference tags to synchronize the readers. Then, TDOA principles and TOA measurements relative to the reference tags and the target tag are used to estimate the location of the target tag. Bekkali et al. [5] collected RSS

J.-S. Pan, S.-M. Chen, N.T. Nguyen (Eds.): ACIIDS 2012, Part I, LNAI 7196, pp. 239–245, 2012.

measure-ments from reference tags to build a probabilistic radio map of the area and then, the Kalman filtering technique is iteratively applied to estimate the target's location.

The goal of this paper is proposes a straightforward and efficient infant location system to reduce the potential risks of the theft and misuse hold. The system can recognize without difficulty different locations of newborn babies which are attached wristband active RFID tags.

2 Infant Location System Architecture

Figure 1 shows a configuration of the intelligent infant location system based on active RFID. The main purpose of the RFID newborn location system is to achieve real-time tracking newborn position. In order to build the network systems combined RFID and wireless Wi-Fi systems in the whole neonatal hospital, we chose the RFID readers and tags with the frequency of 2.45GHz.

Every RFID tags will transmit the various intensity of electromagnetic wave. The RFID Reader will send the various information packets of every RFID tags through the network to the back-end Server to filter. However, the information received from RFID Reader is not all available, which mingles lots of incomplete signals and noise, etc. Therefore, the information packets should be filtered to remove unnecessary records. According to collected information packets, filter server will filter out the signal which is too low, loses some data or is invalid. The remaining information packets will be saved in database. The remaining information packets will be used by the location classifier as experimental data in order to recognize the locations of active RFID tags. The feature and specification of active RFID reader are summarized as table1. The feature and specification of active RFID tag are summarized as table2.

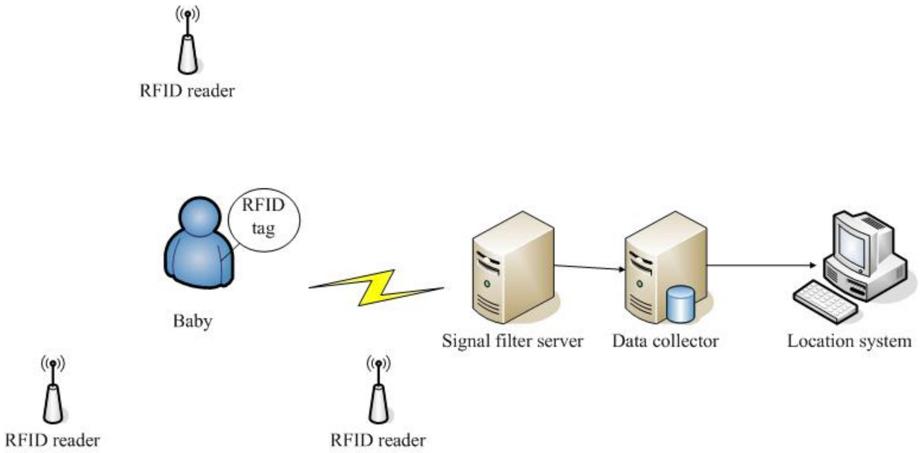


Fig. 1. System Architecture of Infant Location System

Table 1. Feature and Specification of active rfid reader

Feature	
Support 316 RF channels	
Support anti-collision	
Support buzzer alarm	
LED Multi-LED visual indication	
Baud Rate 2,400 bps ~ 115,200 bps	
UID: Tag's identification number	
Support RSSI values 0-255 and RSSI values are inverse proportion	
Specification	
Frequency	2.45GHz
Modulation	FSK
Distance	100 m
Power Consumption	3.0 mm
Transmission power	10dBm
Receiver sensitivity	-103dBm
Iinterface	Support USB / RS232 / RS422 / RS485 / TCPIP
Dimension	107W x 138H x 30D (mm)
Software	Provide WinXP / VISTA / Win7 / WinCE / Linux SDK library for software development.

Table 2. Feature and Specification of active rfid tag

Feature
Wristband design.
Call button: Emergency reporting / Signal transmission.
Remote ON/OFF Tag.
Wireless tag programming.
Two colors LED visual indication: Generally, the emitting signal will glitter green; when it's low battery or detect light sensor, it will glitter red sight.
Built-in light sensor for tamper proof.
Buzzer：Remote active beep or click active beep.

Specification	
Frequency	2.45GHz
Modulation	FSK
Distance	100 m
Power Consumption	3.0 mm
Transmission power	10dBm
Receiver sensitivity	-103dBm
Battery life	3 years(when the transmission number is 10 for each day)，5 years in standby mode.

The remaining information packets include too many information such as index of RFID tag, index of RFID reader, date and time, signal intensity, etc. However, only the signal intensities are considered in this study to recognize the locations of active RFID tags. The signal intensities are transformed to RSSI values and the RSSI values are saved in database.

3 Location Algorithm

Because the signal of RFID is susceptible to environmental factors such as signal refraction, scattering, multi-path effects, etc in indoor environments, it causes the signal strength RSSI values received by the RFID readers change up or down. Moreover, sometimes those medical equipments may have a strong blocking effect to radio signals. These uncertain environmental factors can lead to the accuracy of the neonatal positioning system is not good.

To recognize the locations of active RFID tags, a machine learning techniques are used. In this study, the decision tree-based classifier is developed to decide the correct RFID tags position.

ID3 decision tree algorithm is one of the earliest use, whose main core is to use a recursive form to cut training data. In each time generating node, some subsets of the training input tests will be drawn out to obtain the volume of information coming as a test. After selection, it will yield the greatest amount of value of information obtained as a branch node, selecting the next branch node in accordance with its recursively moves until the training data for each part of a classification fall into one category or meet a condition of satisfaction. C4.5 is the ID3 extension of the method which improved the ID3 excessive subset that contains only a small number of data issues, with handling continuous values-based property, noise processing, and having both pruning tree ability. C4.5 decision tree in each node use information obtained on the volume to select test attribute, to select the information obtained with the highest volume (or maximum entropy compression) of the property as the current test attribute node.

Let A be an attribute with k outcomes that partition the training set S into k subsets S_j (j = 1,..., k). Suppose there are m classes, denoted $C=\{c_1, \quad \cdots \quad ,c_m\}$, and $p_i=\frac{n_i}{n}$ represents the proportion of instances in S belonging to class c_i, where $n=|S|$ and n_i is the number of instances in S belonging to c_i. The selection measure relative to data set S is defined by:

$$Info(S)=\sum_{i=1}^{m} p_i \log_2 p_i \tag{1}$$

The information measure after considering the partition of S obtained by taking into account the k outcomes of an attribute A is given by:

$$Info(S,A)=\sum_{j=1}^{k}\frac{|S_j|}{|S|}Info(S_i) \tag{2}$$

The information gain for an attribute A relative to the training set S is defined as follows:

$$Gain(S,A)=Info(S)-Info(S,A) \tag{3}$$

The Gain(S, A) is called attribute selection criterion. It computes the difference between the entropies before and after the partition, the largest difference corresponds to the best attribute. Information gain has the drawback to favour attributes with a large number of attribute values over those with a small number. To avoid this drawback, the information gain is replaced by a ratio called gain ratio:

$$GR(S,A) = \frac{\mathrm{Gain(S,A)}}{-\sum_{\mathrm{j=1}}^{\mathrm{k}} \frac{|\mathrm{S_i}|}{|S|} \log_2 \frac{|S_i|}{|S|}} \tag{4}$$

Consequently, the largest gain ratio corresponds to the best attribute.

In this study, C5.0 is adapted to generate decision tree and to produce the result. C4.5 algorithm is a continuation framework. Unlike C4.5, C5.0 provide orders which are the most popular set of rules in many applications. The classification conditions expressed in the form of rules one by one, which increase the classified readability. C5.0 can deal with the future or nominal values of the various of information. The results can easily be understood. The difference between C5.0 and C4.5 is that C5.0 can handle many more data types such as date, time, time stamp, sequence type of discrete data, etc.

4 Experimental Environment

The equipment used in this study included three RFID readers, six newborn babies attached wristband active RFID tags, for each connected network devices to the filter server of signal packet. Experimental environment is 3.0 meters * 3.0 meters as shown in Figure 2. Figure 2 also shows that experimental region is divided into nine regional grids.

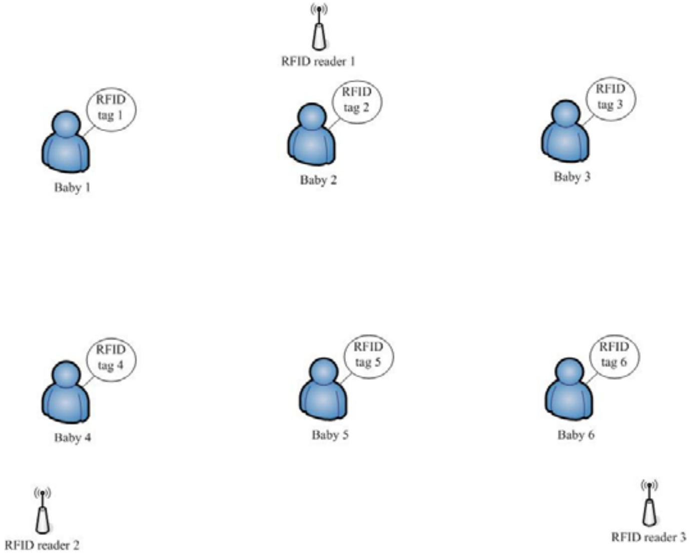


Fig. 2. Experimental environment

5 Numerical Results

Every active RFID tags send an electromagnetic wave per second. The intensity of electromagnetic waves transmitted by active RFID tags is transformed to the RSSI value. Each reader support RSSI values 0-255. The reading range and RSSI are inverse proportion. The number of experimental records is 300, including 75% of experimental records as training data and 25% of experimental records as training data, in this study. The experimental accuracies of RFID Tags locations are shown in Table 3.

Table 3. Experimental Accuracy

	Baby 1	*Baby 2*	*Baby 3*	*Baby 4*	*Baby 5*	*Baby 6*
Number of training records	225	225	225	225	225	225
Number of test records	75	75	75	75	75	75
Accuracy in training phase	100%	100%	100%	100%	100%	100%
Accuracy in test phase	99%	100%	100%	100%	100%	100%

6 Conclusion

This paper presents an intelligent infrant location system based on active RFID in conjunction with decision tree-based classifier. From the experimental results obtained in this study, some conclusions can be summarized as follows:

(1.) The experimental results show that the proposed infrant location system can accurately recognizes the locations of newborn babies.

(2.) The method presented in the study is straightforward, simple and valuable for practical applications.

Acknowledgements. This work was partially supported by the National Science Council, Republic of China, under Grant NSC 99-2815-C-270-010-H. The author also thanks the Chienkuo Technology University (Contract No. CTU-99-RP-IM-001-020-A) for financial supports.

References

1. Hightower, J., Vakili, C., Borriello, C., Want, R.: Design and Calibration of the SpotON AD-Hoc Location Sensing System, UWCSE 00-02-02 University of Washington, Department of Computer Science and Engineering, Seattle, `http://www.cs.washington.edu/homes/jeffro/pubs/hightower2001design/hightower2001design.pdf`

2. Ni, L.M., Liu, Y., Lau, Y.C., Patil, A.P.: LANDMARC:Indoor Location Sensing Using Active RFID (2003)
3. Wang, C., Wu, H., Tzeng, N.F.: RFID-based 3-D positioning schemes. In: IEEE INFOCOM, pp. 1235–1243 (2007)
4. Stelzer, A., Pourvoyeur, K., Fischer, A.: Concept and application of LPM—a novel 3-D local position measurement system. IEEE Trans. Microwave Theory Techniques 52(12), 2664–2669 (2004), http://www.ubisense.net/default.aspS
5. Bekkali, A., Sanson, H., Matsumoto, M.: RFID indoor positioning based on probabilistic RFID map and kalman filtering. In: 3rd IEEE International Conference on Wireless and Mobile Computing, Networking and Communications, IEEE WiMob (2007)

An Intelligently Remote Infant Monitoring System Based on RFID

Shou-Hsiung Cheng

Department of Information Management,
Chienkuo Technology University, Changhua 500, Taiwan
shcheng@ctu.edu.tw

Abstract. This study proposes a straightforward and efficient intelligently remote infant monitoring system to reduce the potential risks of the theft, misuse hold and abnormal body temperature. The system can accurately recognizes the locations of newborn babies by using neural network classifiers after the active RFID readers has received different intensity of electromagnetic waves transmitted by active RFID tags. The newborn babies of temperature anomalies also can be diagnosed by the body temperature sensors and the proposed infant monitoring system. The remote infant monitoring system improved infant care and safety, reduced systems and human-based errors and enabled fast communicating with the clinical staff and families. This system can be used for infants at home or in a hospital nursery room.

Keywords: infant monitoring system, RFID, neural networks classifiers.

1 Introduction

As the characteristics of newborn babies are similar and the lack of expression ability, it often lead to neonatal injury due to human negligence or error identified. Infancy is the most vulnerable stage, a variety of injuries can cause permanent sequelae or death. How to enhance the identification of newborns and to prevent misuse hold are the important projects in neonatal care. Infant hearthcare with RFID discarding the traditional paper job is becoming more and more urgent need as hospitals in today's competitive environment.

A highly concerning healthcare application is real-time tracking and location of medical assets. Swedberg [1] proposed a patient-and employee-tracking system based on active 433MHz RFID technology is currently being tested at Massachusetts General Hospital. The pilot gathers information regarding patient flow and bottlenecks with the expected outcome of gaining a better understanding of how the clinical system behaves. It will potentially reveal aspects such as how long a patient sat alone in an examining room or whether the medical personnel spent the proper time with the patient. Chowdhury et al. [2] were patients are assigned a RFID wristband for identification purposes, but their approach does not take advantage of storage capabilities of RFID devices. The unique identification code provided by the wristband is only used as a ''license plate'' and all related data is stored and recovered in a backend server as in traditional non-RFID patient management systems.

J.-S. Pan, S.-M. Chen, N.T. Nguyen (Eds.): ACIIDS 2012, Part I, LNAI 7196, pp. 246–254, 2012.

In recent years, Radio Frequency Identification (RFID) technology has been widely accepted in hospitals to track and locate newborn babies. RFID positioning systems can be broadly divided into two classes: tag and reader localization, depending on the RFID component type of the target. In tag localization schemes, readers and possibly tags are deployed as reference points within the area of interest and a positioning technique is applied for estimating the location of a tag. Hightower [3] was the first to use RFID technology to build the indoor positioning system. He uses RSS measurements and proposed SpotON system to estimate the distance between a target tag and at least three readers and then applies trilateration on the estimated distances. Lionel [4] proposed the LANDMARC method based on RFID to build indoor positioning system. LANDMARC system have a greater read range and response capability of the active sensor tags compared to SpotON. Wang et al. [5] propose a 3-D positioning scheme, namely passive scheme, which relies on a deployment of tags and readers with different power levels on the floor and the ceiling of an indoor space and uses the Simplex optimization algorithm for estimating the location of multiple tags. On the other hand, in tag localization schemes, usually passive or active tags with known coordinates and possibly readers are deployed as reference points and their IDs are associated with their location information. Lee et al. [6] proposed passive tags are arranged on the floor at known locations in square pattern. The reader acquires all readable tag locations and estimates its location and orientation by using weighted average method and Hough transform, respectively. Yamano et al. [7] utilize the received signal strength to determine the reader's position by using machine learning technique. In the training phase, the reader acquires the RSS from every tag in various locations in order to build a support vector machine. Xu et al. [8] proposed a Bayesian approach to predict the position of a moving object.

This study proposes a straightforward and efficient remote infant monitoring system to reduce the potential risks of the theft, misuse hold and abnormal body temperature. Not only the proposed monitoring location system can recognize different babies but also can track the locations of newborn babies by using active RFID tags. Further, the proposed infant monitoring system can send off warning signals when the theft, misuse hold and abnormal body temperature of the babies are occurred. The remote infant monitoring system enable fast communicating with the clinical staff and families by the mobile devices such as notebook, tablet PC, and smart phone.

2 Intelligently Remote Infant Monitoring System Architecture

Figure 1 shows a configuration of the intelligent infant location system based on active RFID. In order to care the newborn, the newborn in the baby room will be attached wristband active RFID tags as shown in Figure 2 by the nurses. The main purpose of the RFID newborn positioning system is to achieve real-time tracking newborn position and newborn body temperature. Through this newborn monitoring system, doctors, nurses and parents of newborn babies can always check the location to prevent the risk of theft, misuse hold neonatal and abnormal body temperature by the mobile devices such as notebook, tablet PC, and smart phone. When abnormal conditions of newborns are occurred, the nursing staff will control the abnormal situation immediately and they will handle responsibly after receipt of warning.

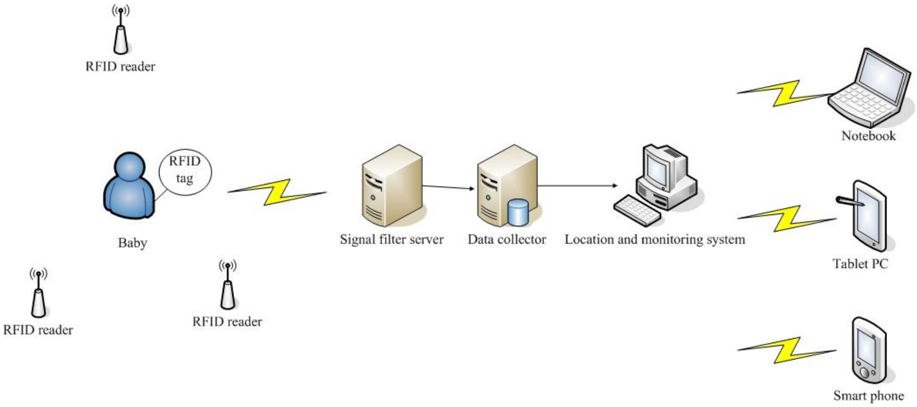


Fig. 1. System Architecture of the Intelligent Remote Infant Monitoring System

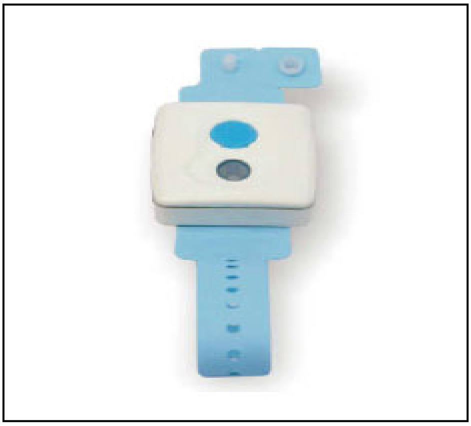


Fig. 2. Wristband Active RFID Tags

In order to build the network systems combined RFID and wireless Wi-Fi systems in the whole neonatal hospital, we chose the RFID readers and tags with the frequency of 2.45GHz. Every RFID tags will transmit the various intensity of electromagnetic wave. The RFID Reader will send the various information packets of every RFID tags through the network to the back-end Server to filter. However, the information received from RFID Reader is not all available, which mingles lots of incomplete signals and noise, etc. Therefore, the information packets should be filtered to remove unnecessary records. According to collected information packets, filter server will filter out the signal which is too low, loses some data or is invalid. The remaining information packets will be saved in database. The remaining information packets will be used by the location classifier as experimental data in order to recognize the locations of active RFID tags. The feature and specification of active RFID reader are summarized as table1. The feature and specification of active RFID tag are summarized as table2. The wristband active RFID tag are built in two thermal sensors for continuously monitoring the temperature of infants. One thermal sensor detects the skin temperature. The other detects the ambient temperature.

Table 1. Feature and Specification of active rfid reader

Feature
Support 316 RF channels
Support anti-collision
Support buzzer alarm
LED Multi-LED visual indication
Baud Rate 2,400 bps ~ 115,200 bps
UID: Tag's identification number
T1: Ambient temperature sensor
T2: Skin temperature sensor
Note: T1 / T2 / SENSOR use for anti-tamper capability.
Support RSSI values 0-255 and RSSI values are inverse proportion
RSSI: Received Signal Strength Indication (0-255). Reading range and RSSI are inverse proportion.

Specification	
Frequency	2.45GHz
Modulation	FSK
Distance	100 m
Power Consumption	3.0 mm
Transmission power	10dBm
Receiver sensitivity	-103dBm
Iinterface	Support USB / RS232 / RS422 / RS485 / TCPIP
Dimension	107W x 138H x 30D (mm)
Software	Provide WinXP / VISTA / Win7 / WinCE / Linux SDK library for software development.

Table 2. Feature and Specification of active rfid tag

Feature
Wristband design.
Call button: Emergency reporting / Signal transmission.
Remote ON/OFF Tag.
Wireless tag programming.
Two colors LED visual indication: Generally, the emitting signal will glitter green; when it's low battery or detect light sensor, it will glitter red sight.
Built-in two thermal sensors for continuously monitoring the temperature of infants. One thermal sensor detects the skin temperature. The other detects the ambient temperature.
Built-in light sensor for tamper proof.
Buzzer：Remote active beep or click active beep.

Specification	
Frequency	2.45GHz
Modulation	FSK
Distance	100 m
Power Consumption	3.0 mm
Transmission power	10dBm
Receiver sensitivity	-103dBm
Battery life	3 years(when the transmission number is 10 for each day)，5 years in standby mode.

The remaining information packets include too many information such as index of RFID tag, index of RFID reader, date and time, signal intensity, ambient temperature , and skin temperature etc. However, only the signal intensities are considered in this study to recognize the locations of active RFID tags. The signal intensities are transformed to RSSI values and the RSSI values are saved in database. The ambient temperature and skin temperature are collected and saved in database. The proposed remote infant monitoring system can send off warning signals when the theft, misuse hold and abnormal body temperature of the babies are occurred. The remote infant monitoring system enable fast communicating with the clinical staff and families by the mobile devices such as notebook, tablet PC, and smart phone.

3 Location Algorithm

Because the signal of RFID is susceptible to environmental factors such as signal refraction, scattering, multi-path effects, etc in indoor environments, it causes the signal strength RSSI values received by the RFID readers change up or down. Moreover, sometimes those medical equipments may have a strong blocking effect to radio signals. These uncertain environmental factors can lead to the accuracy of the neonatal positioning system is not good.

To recognize the locations of active RFID tags, a machine learning techniques are used. In this study, the neural network classifier is developed to decide the correct RFID tags position. The basic element of a artificial neural network is a neuron. This is a simple virtual device that accepts many inputs, sums them, applies a nonlinear transfer function, and generates the result, either as a model prediction or as input to other neurons. A neural network is a structure of many such neurons connected in a systematic way. The neurons in such networks are arranged in layers. Normally, there is one layer for input neurons, one or more layers of the hidden layers, and one layer for output neurons. Each layer is fully interconnected to the preceding layer and the following layer. The neurons are connected by links, each link has a numerical weight associated with it, which determine the weights are the basic means of long-term memory in artificial neural network. A neural networks learns through repeated adjustments of these weights.

In this study, a radial basis function network is selected. It consists of three layers: an input layer, a receptor layer, and an output layer. The input and output layers are similar to those of a multilayer perceptron. However, the hidden or receptor layer consists of neurons that represent clusters of input patterns, similar to the clusters in a k-means model. These clusters are based on radial basis functions. The connections between the input neurons and the receptor weights are trained in essentially the same manner as a k-means model. Particularly, the receptor weights are trained with only the input fields; the output fields are ignored for the first phase of training. Only after the receptor weights are optimized to find clusters in the input data are the connections between the receptors and the output neurons trained to generate predictions. Each receptor neuron has a radial basis function associated with it. The basis function used in the study is a multidimensional Gaussian function,

$$\exp \quad \frac{d_i^2}{2\sigma_i^2}$$

where d_i is the distance from cluster center i, and σ_i is a scale parameter describing the size of the radial basis function for cluster i.

The scale parameter σ_i is calculated based on the distances of the two closest clusters, $\sqrt{\frac{d_1 + d_2}{2}}$, where d_1 is the distance between the cluster center and the center of the closest other cluster, and d_2 is the distance to the next closest cluster center. Thus, clusters that are close to other clusters will have a small receptive field, while those that are far from other clusters will have a larger receptive field.

During output weight training, records are presented to the network as with a multilayer perceptron. The receptor neurons compute their activation as a function of their radial basis function size and the user-specified overlapping value h. The activation for receptor neuron j is calculated as

$$\sigma_j = e^{\frac{\|r \;\; c\|^2}{2\sigma_j^2 h}} \tag{1}$$

where r is the vector of record inputs and c is the cluster center vector. The output neurons are fully interconnected with the receptor or hidden neurons. The receptor neurons pass on their activation values, which are weighted and summed by the output neuron,

$$O_k = \sum_j W_{jk} a_j \tag{2}$$

The output weights W_{jk} are trained in a manner similar to the training of a two-layer back-propagation network. The weights are initialized to small random values in the range $-0.001 \le w_{ij} \le 0.001$, and then they are updated at each cycle p by the formula

$$w_{jk}(p) = w_{jk}(p-1) + \Delta w_{jk}(p) \tag{3}$$

The change value is calculated as

$$\Delta w_{jk}(p) = \eta(r_k - O_k)a_i + \alpha \Delta w_{jk}(p-1) \tag{4}$$

which is analogous to the formula used in the back-propagation method.

4 Experimental Environment

The equipment used in this study included three RFID readers, nine active RFID tags with same specification and feature, for each connected network devices to the filter server of signal packet. Experimental environment is 3.0 meters * 3.0 meters as shown in Figure 2. Figure 2 also shows that experimental region is divided into four regional grids.

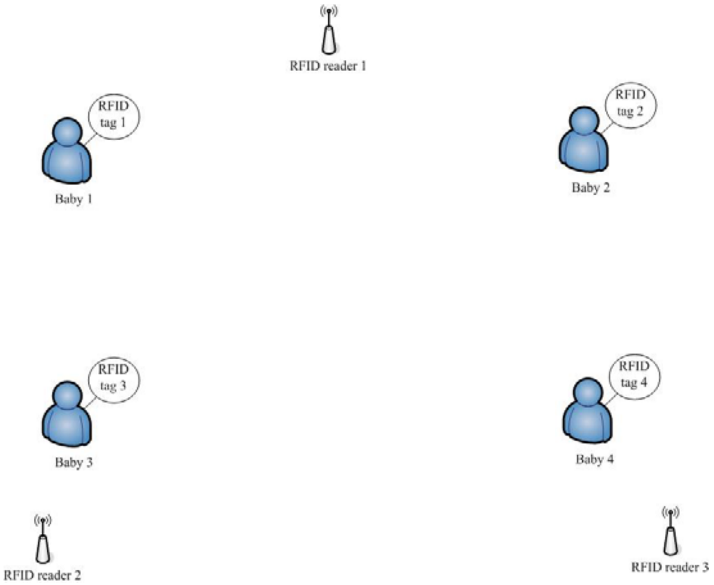


Fig. 3. Experimental environment

5 Numerical Results

Every active RFID tags send an electromagnetic wave per second. The intensity of electromagnetic waves transmitted by active RFID tags is transformed to the RSSI value. Each reader support RSSI values 0-255. The reading range and RSSI are inverse proportion. The number of experimental records is 500, including 75% of experimental records as training data and 25% of experimental records as training data, in this study. The experimental accuracies of RFID Tags locations are shown in Table 3. The Screen of intelligently remote infant monitoring system is shown in Figure 4.

Table 3. Experimental Accuracy

	Baby 1	*Baby 2*	*Baby 3*	*Baby 4*
Number of training records	375	375	375	375
Number of test records	125	125	125	125
Accuracy in training phase	100%	100%	100%	100%
Accuracy in test phase	100%	100%	100%	100%

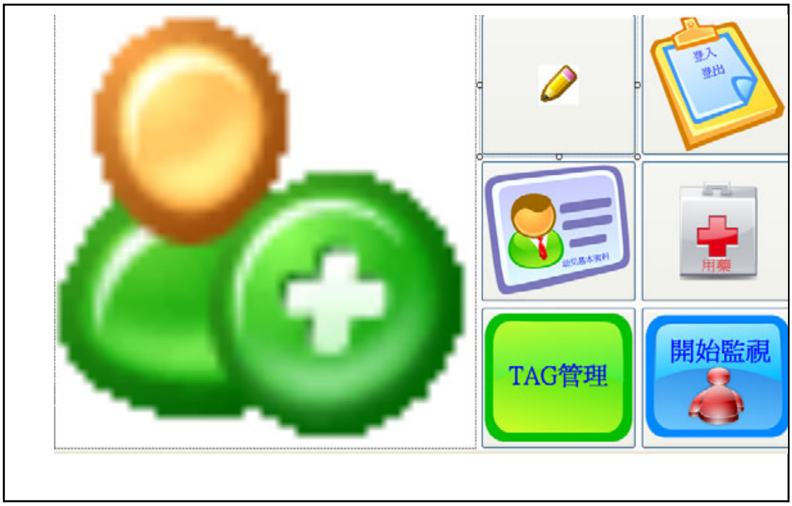


Fig. 4. The Screen of Intelligently Remote Infant Monitoring System

This paper presents an intelligently remote infant monitoring system based on active RFID in conjunction with neural networks. From the experimental results obtained in this study, some conclusions can be summarized as follows:

(1.) The experimental results show that the proposed remote infant monitoring system can accurately recognizes the locations of newborn babies.

(2.) The remote infant monitoring system presented in the study is straightforward, simple and valuable for practical applications.

Acknowledgements. This work was partially supported by the National Science Council, Republic of China, under Grant NSC 99-2815-C-270-010-H. The author also thanks the Chienkuo Technology University (Contract No. CTU-99-RP-IM-001-020-A) for financial supports.

References

1. Chowdhury, B., Khosla, R.: RFID-based hospital real-time patient management system. In: 6th IEEE/ACIS International Conference on Computer and Information Science, ICIS 2007, pp. 363–368 (2007)
2. Swedberg, C.: Massachusetts general uses RFID to better understand its clinics (October 2009c), `http://www.rfidjournal.com/article/view/5324/S`
3. Hightower, J., Vakili, C., Borriello, C., Want, R.: Design and Calibration of the SpotON AD-Hoc Location Sensing System, UWCSE 00-02-02 University of Washington, Department of Computer Science and Engineering, Seattle, `http://www.cs.washington.edu/homes/jeffro/pubs/hightower2001design/hightower2001design.pdf`

4. Ni, L.M., Liu, Y., Lau, Y.C., Patil, A.P.: LANDMARC:Indoor Location Sensing Using Active RFID (2003)
5. Wang, C., Wu, H., Tzeng, N.: RFID-based 3-D positioning schemes. In: IEEE INFOCOM, pp. 1235–1243 (2007)
6. Lee, H.J., Lee, M.: Localization of mobile robot based on radio frequency identification devices. In: SICE-ICASE, International Joint Conference, pp. 5934–5939 (October 2006)
7. Yamano, K., et al.: Self-localization of mobile robots with RFID system by using support vector machine. In: Proceedings of 2004 IEEWRSI International Conference on Intelligent Robots and Systems, Sendai, Japan (2004)
8. Xu, B., Gang, W.: Random sampling algorithm in RFID indoor location system. In: IEEE International Workshop on Electronic Design, Test and Applications, DELTA 2006 (2006)

An Intelligent Infant Monitoring System Using Active RFID

Shou-Hsiung Cheng

Department of Information Management,
Chienkuo Technology University, Changhua 500, Taiwan
shcheng@ctu.edu.tw

Abstract. At present, Radio Frequency Identification (RFID) technology has been widely accepted in hospitals to monitoring newborn babies.This study proposes a straightforward and efficient infant monitoring system to reduce the potential risks of the theft, misuse hold and abnormal body temperature. The system can recognize without difficulty different locations of newborn babies which are attached wristband active RFID tags. The system can accurately recognizes the locations of newborn babies by using Bayesian network classifier after the active RFID readers has received different intensity of electromagnetic waves transmitted by active RFID tags. The infant monitoring system also can detect temperature anomalies of newborn babies real time by the body temperature sensors.

Keywords: infant monitoring system, RFID, Bayesian network classifier.

1 Introduction

Because newborn babies are often difficult to distinguish and the ability of expression is lack, it often cause of neonatal injury due to human negligence or error identified. Infancy is the most vulnerable stage, a variety of injuries can cause permanent sequelae or death. The staff is complex and the flow of people is very large in neonatal hospital. How to enhance the identification of newborns and to prevent misuse hold and to monitor the physiological status of newborn babies are the important projects in neonatal care. It is becoming more and more important at hospitals to establish an infant monitoring system using RFID in today's competitive environment.

In recent years, high expectations for the integration of RFID in healthcare scenaios have emerged. However, In spite of recent reasearch interest in the healthcare enviroment, RFID adoption is still in its infancy and a larger number of experience need to be collected. As a consequence of Severe Acute Respiratory Syndrome(SARS) where 37 patients died and part of the medical personnel was also infected the Ministry of Economic Affairsin Taiwan granted research funds to support the implementation of RFID in healthcare. Tzeng et al. [1] has been presented the experience of five early adopters hospitals. Authors conclude that future empirical research will be helpful invalidating their propositions requiring a bigger number of

J.-S. Pan, S.-M. Chen, N.T. Nguyen (Eds.): ACIIDS 2012, Part I, LNAI 7196, pp. 255–263, 2012.

experiences collected and studied. However they consider RFID useful in enhancing patient care and analyzing workload of medical staff. Holzinger et al. [2] proposed an experience of tracking elderly patients suffering from dementia was also tested in the Albert Schweitzer II Hospital with the purpose of providing real-time location and an alert system if a patient goes beyond his expected location. Medical personnel provided positive feedback, but patients themselves reacted negatively to the idea of surveillance.

At present, Radio Frequency Identification (RFID) technology has been widely accepted in hospitals to track and locate newborn babies. RFID positioning systems can be broadly divided into two classes: tag and reader localization, depending on the RFID component type of the target. In tag localization schemes, readers and possibly tags are deployed as reference points within the area of interest and a positioning technique is applied for estimating the location of a tag. Hightower [3] was the first to use RFID technology to build the indoor positioning system. He uses RSS measurements and proposed SpotON system to estimate the distance between a target tag and at least three readers and then applies trilateration on the estimated distances. Lionel [4] proposed the LANDMARC method based on RFID to build indoor positioning system. LANDMARC system have a greater read range and response capability of the active sensor tags compared to SpotON. On the other hand, in tag localization schemes, usually passive or active tags with known coordinates and possibly readers are deployed as reference points and their IDs are associated with their location information. Lee et al. [5] proposed passive tags are arranged on the floor at known locations in square pattern. The reader acquires all readable tag locations and estimates its location and orientation by using weighted average method and Hough transform, respectively. Han et al. [6] arrange tags in triangular pattern so that the distance in x-direction is reduced. They show that the maximum estimation error is reduced about 18% from the error inthe square pattern.

This study proposes a straightforward and efficient infant monitoring system to reduce the potential risks of the theft, misuse hold and abnormal body temperature. Not only the proposed monitoring location system can recognize different babies but also can track the locations of newborn babies by using active RFID tags. Further, the proposed infant monitoring system can send off warning signals when the theft, misuse hold and abnormal body temperature of the babies are occurred.

2 The Architecture of the Intelligent Infant Monitoring System

Figure 1 shows a configuration of the intelligent infant location system based on active RFID. In order to care the newborn, the newborn in the baby room will be attached wristband active RFID tags as shown in Figure 2 by the nurses. The main purpose of the RFID newborn monitoring system is to achieve real-time tracking newborn position and newborn body temperature.

Wristband active RFID tags can be issued to every newborn babies at registration, and then it can be used to identify patients during the entire hospitalization period. It can also be used to store newborn baby important data (such as name, patient ID, drug allergies, drugs that the patient is on today, blood group, and so on) in order to dynamically inform staff before critical situation. The encoded data of wristband

RFID tags can be read through bed linens, while newborn babies are sleeping without disturbing them. RFID technology provides a method to transmit and receive data from a newborn baby to health service provider/medical professionals without human intervention (i.e., wireless communication). It is an automated data-capture technology that can be used to identify, track, and store patient information electronically contained on RFID wristband smart tag. Although, medical professionals/consultants can access/update patient's record remotely via Wi-Fi connection using mobile devices such as Personal Digital Assistant (PDA), laptops and other mobile devices. Wi-Fi, the wireless local area networks (WLAN) that allows healthcare provider (e.g., hospitals) to deploy a network more quickly, at lower cost, and with greater flexibility than a wired system.

Every RFID tags will transmit the various intensity of electromagnetic wave. The RFID Reader will send the various information packets of every RFID tags through the network to the back-end Server to filter. However, the information received from RFID Reader is not all available, which mingles lots of incomplete signals and noise, etc. Therefore, the information packets should be filtered to remove unnecessary records. According to collected information packets, filter server will filter out the signal which is too low, loses some data or is invalid. The remaining information packets will be saved in database. The remaining information packets will be used by the location classifier as experimental data in order to recognize the locations of active RFID tags. When abnormal conditions of newborns are occurred, the nursing staff will control the abnormal situation immediately and they will handle responsibly after receipt of warning.

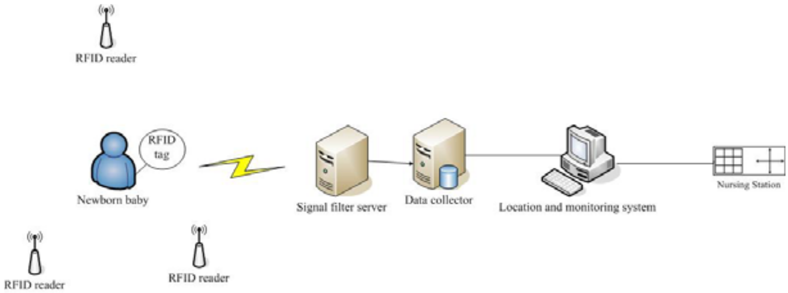


Fig. 1. System Architecture of Intelligent Infant Monitoring System

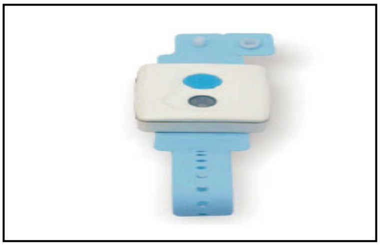


Fig. 2. Wristband Active RFID Tags

Table 1. Feature and Specification of active rfid reader

Feature
Support 316 RF channels
Support anti-collision
Support buzzer alarm
LED Multi-LED visual indication
Baud Rate 2,400 bps ~ 115,200 bps
UID: Tag's identification number
T1: Ambient temperature sensor
T2: Skin temperature sensor
Note: T1 / T2 / SENSOR use for anti-tamper capability.
Support RSSI values 0-255 and RSSI values are inverse proportion
RSSI: Received Signal Strength Indication (0-255). Reading range and RSSI are inverse proportion.

Specification	
Frequency	2.45GHz
Modulation	FSK
Distance	100 m
Power Consumption	3.0 mm
Transmission power	10dBm
Receiver sensitivity	-103dBm
Iinterface	Support USB / RS232 / RS422 / RS485 / TCPIP
Dimension	107W x 138H x 30D (mm)
Software	Provide WinXP / VISTA / Win7 / WinCE / Linux SDK library for software development.

Table 2. Feature and Specification of active rfid tag

Feature
Wristband design.
Call button: Emergency reporting / Signal transmission.
Remote ON/OFF Tag.
Wireless tag programming.
Two colors LED visual indication: Generally, the emitting signal will glitter green; when it's low battery or detect light sensor, it will glitter red sight.
Built-in two thermal sensors for continuously monitoring the temperature of infants. One thermal sensor detects the skin temperature. The other detects the ambient temperature.
Built-in light sensor for tamper proof.
Buzzer：Remote active beep or click active beep.

Specification	
Frequency	2.45GHz
Modulation	FSK
Distance	100 m
Power Consumption	3.0 mm
Transmission power	10dBm
Receiver sensitivity	-103dBm
Battery life	3 years(when the transmission number is 10 for each day)，5 years in standby mode.

The remaining information packets include too many information such as index of RFID tag, index of RFID reader, date and time, signal intensity, ambient temperature, and skin temperature etc. However, only the signal intensities are considered in this study to recognize the locations of active RFID tags. The signal intensities are transformed to RSSI values and the RSSI values are saved in database. The ambient temperature and skin temperature are collected and saved in database. The proposed infant monitoring system can send off warning signals to nursing station when the theft, misuse hold and abnormal body temperature of the babies are occurred. The infant monitoring system also can detect temperature anomalies of newborn babies real time by the body temperature sensors.

3 The Algorithms of Positioning Newborn Babies

Because the signal of RFID is susceptible to environmental factors such as signal refraction, scattering, multi-path effects, etc in indoor environments, it causes the signal strength RSSI values received by the RFID readers change up or down. Moreover, sometimes those medical equipments may have a strong blocking effect to radio signals. These uncertain environmental factors can lead to the accuracy of the neonatal positioning system is not good.

To recognize the locations of active RFID tags, a machine learning techniques are used. In this study, the Naïve Bayesian network classifier is considered to decide the correct RFID tags position. The tree augmented Naïve Bayesian networks algorithm is considered to classify the position of active RFID tags in this study. This algorithm creates a simple Bayesian network model. This model is an improvement over the naïve Bayesian model as it allows for each predictor to depend on another predictor in addition to the target attribute. Its main advantages are its classification accuracy and favorable performance compared with general Bayesian network models. Not only it is simply but also it imposes much restriction on the dependency structure uncovered among its nodes. The Bayesian network classifier is a simple classification method, which classifies a case

$$d_j = \left(x_1^j, x_2^j, \cdots, x_n^j\right)$$

by determining the probability. The probabilities are calculated as

$$P_r = \left(Y_i \mid X_1 = x_1^j, X_2 = x_2^j, ..., Xn = x_n^j\right)$$

Equation (1) also can be rewritten as

$$P_r = \frac{P_r(Y_i)P_r\left(X_1 = x_1^j, X_2 = x_2^j, ..., X_n = x_n^j \mid Y_i\right)}{P_r\left(X_1 = x_1^j, X_2 = x_2^j, ..., X_n = x_n^j\right)}$$

$$P_r \propto P_r(Y_i)\prod_{k=1}^{n} P_r\left(X_k = x_k^j \mid \pi_k^j, Y_i\right)$$

TAN Structure Learning

We use a maximum weighted spanning tree method to construct a tree Bayesian network. This method associates a weight to each edge corresponding to the mutual information between the two variables. When the weight matrix is created, the algorithm gives an undirected tree that can be oriented with the choice of a root.

The mutual information of two nodes is defined as

$$I\left(X_i, X_j\right) = \sum_{x_i, x_j} P_r\left(x_i, x_j\right) \log\left(\frac{P_r\left(x_i, x_j\right)}{P_r\left(x_i\right) P_r\left(x_j\right)}\right)$$

We replace the mutual information between two predictors with the conditional mutualinformation between two predictors given the target. It is defined as

$$I\left(X_i, X_j \mid Y\right) = \sum_{x_i x_j y_k} P_r\left(x_i, x_j, y_k\right) \log\left(\frac{P_r\left(x_i, x_j y_k\right)}{P_r\left(x_i \mid y_k\right) P_r\left(x_j \mid y_k\right)}\right)$$

TAN Parameter Learning

Let be the cardinality of . Let denote the cardinality of the parent set of , that is, the number of different values to which the parent of can be instantiated. So it can be calculated as . Note implies . We use to denote the number of records in D for which takes its jth value. We use to denote the number of records in D for which take its jth value and for which takes its kth value.

Maximum Likelihood Estimation

The closed form solution for the parameters and that maximize the log likelihood score is

$$\overset{\wedge}{\theta} Y_i = \frac{N_{Y_i}}{N}$$

$$\overset{\wedge}{\theta} Y_{ijK} = \frac{N_{ijk}}{N_{ij}}$$

$$K = \sum_{i=1}^{n} \left(r_i - 1\right) \cdot q_r + |Y| - 1$$

TAN Posterior Estimation

Assume that Dirichlet prior distributions are specified for the set of parameters as well as for each of the sets. Let and denote corresponding Dirichlet distribution parameters such that and. Upon observing the dataset D, we obtain Dirichlet posterior distributions with the following sets of parameters:

$$\overset{\wedge}{\theta}{}^{P}_{Y_i} = \frac{N_{Y_i} + N^0_{Y_i}}{N + N^0}$$

$$\overset{\wedge}{\theta}{}^{P}_{ijk} = \frac{N_{ijk} + N^0_{ijk}}{N_{ij} + N^0_{ij}}$$

The posterior estimation is always used for model updating.

4 Experimental Environment

The equipment used in this study included three RFID readers, nine active RFID tags with same specification and feature, for each connected network devices to the filter server of signal packet. Experimental environment is 3.0 meters * 3.0 meters as shown in Figure 2. Figure 2 also shows that experimental region is divided into three regional grids.

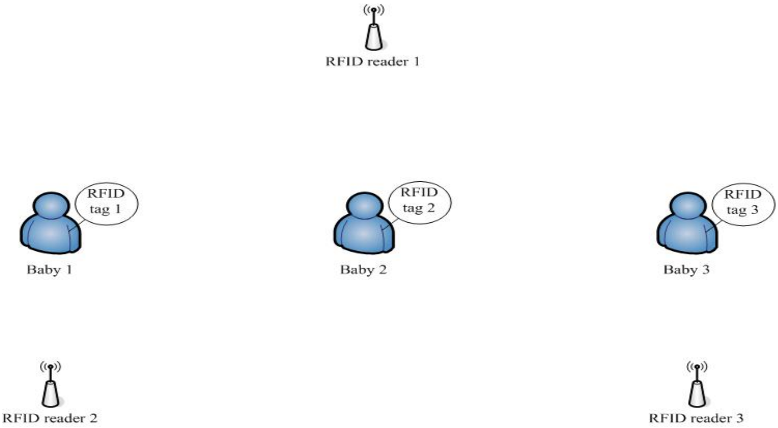


Fig. 3. Experimental environment

5 Numerical Results

Every active RFID tags send an electromagnetic wave per second. The intensity of electromagnetic waves transmitted by active RFID tags is transformed to the RSSI value. Each reader support RSSI values 0-255. The reading range and RSSI are inverse proportion. The number of experimental records is 400, including 75% of experimental records as training data and 25% of experimental records as training data, in this study. The experimental accuracies of RFID Tags locations are shown in Table 3. The screen of intelligent infant monitoring system is shown in Fugure 4.

Table 3. Experimental Accuracy

	Baby 1	*Baby 2*	*Baby 3*
Number of training records	300	300	300
Number of test records	100	100	100
Accuracy in training phase	100%	100%	100%
Accuracy in test phase	100%	100%	100%

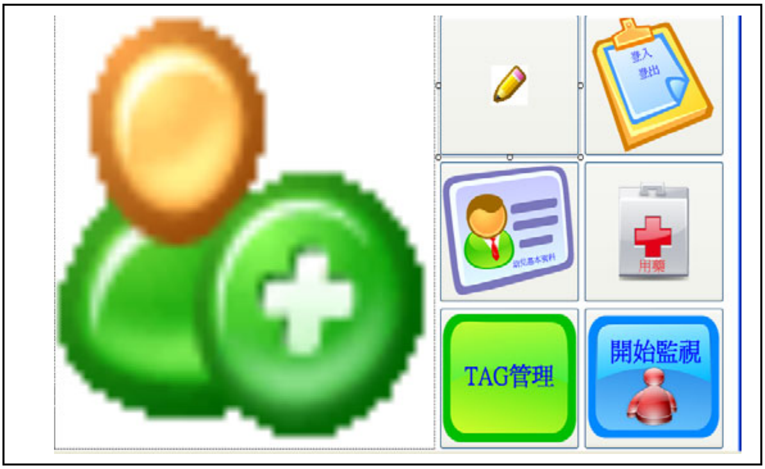


Fig. 4. The Screen of Intelligent Remote Infant Monitoring System

6 Conclusion

This paper presents an intelligent infant monitoring system based on active RFID in conjunction with Bayesian network classifier. From the experimental results obtained in this study, some conclusions can be summarized as follows:

(1.) The experimental results show that the proposed infant monitoring system can accurately recognizes the locations of newborn babies.

(2.) The infant monitoring system also can detect temperature anomalies of newborn babies real time by the body temperature sensors.

(3.) The infant monitoring system presented in the study is straightforward, simple and valuable for practical applications.

References

1. Tzeng, S., Chen, W., Pai, F.: Evaluating the business value of RFID: evidence from five case studies. International Journal of Production Economics, Abr 112(2), 601–613 (2008)
2. Holzinger, A., Schaupp, K., Eder-Halbedl, W.: An Investigation on Acceptance of Ubiquitous Devices for the Elderly in a Geriatric Hospital Environment: Using the Example of Person Tracking. In: Miesenberger, K., Klaus, J., Zagler, W.L., Karshmer, A.I. (eds.) ICCHP 2008. LNCS, vol. 5105, pp. 22–29. Springer, Heidelberg (2008)
3. Hightower, J., Vakili, C., Borriello, C., Want, R.: Design and Calibration of the SpotON AD-Hoc Location Sensing System, UWCSE 00-02-02 University of Washington, Department of Computer Science and Engineering, Seattle, `http://www.cs.washington.edu/homes/jeffro/pubs/hightower2001design/hightower2001design.pdf`

4. Ni, L.M., Liu, Y., Lau, Y.C., Patil, A.P.: LANDMARC:Indoor Location Sensing Using Active RFID (2003)
5. Lee, H.J., Lee, M.: Localization of mobile robot based on radio frequency identification devices. In: SICE-ICASE, International Joint Conference, pp. 5934–5939 (October 2006)
6. Han, S.S., Lim, H.S., Lee, J.: An efficient localization scheme for a differential-driving mobile robot based on RFID system. IEEE Trans. Ind. Electron. 54, 3362–3369 (2007)
7. Hightower, J., Borriello, G.: Location systems for ubiquitous computing. IEEE Computer 34, 57–66 (2001)

Fuzzy Decision Making for Diagnosing Machine Fault

Lily Lin[1], Huey-Ming Lee[2], and Jin-Shieh Su[3]

[1] Department of International Business, China University of Technology,
56, Sec. 3, Hsing-Lung Road, Taipei (116), Taiwan
[2] Department of Information Management, Chinese Culture University
[3] Department of Applied Mathematics, Chinese Culture University
55, Hwa-Kung Road, Yang-Ming-San, Taipei (11114), Taiwan
lily@cute.edu.tw,
{hmlee,sjs}@faculty.pccu.edu.tw

Abstract. The purpose of this study is to present a fuzzy diagnosing machine fault to support the developing machine diagnosis system. The fuzzy evaluation is used to process the problems of which the fault causes and the symptoms are dealing with the uncertainty environment. In this study, we propose two propositions to treat the machine diagnosis fault.

Keywords: Fuzzy inference, Fault diagnosis, Membership grade.

1 Introduction

Though the mature manufacturing technology of machine is developed, to handle machine fault states in real time is still very important. If machines give the fault, manufacturers have to spend a lot of time to find and eliminate the breakdown. If there is an accurate machine fault diagnosis system, the decision makers may improve ability of production and reduce defective rate.

Machine fault diagnosis is not only a traditional maintenance problem but also is an important issue. There are many papers investigating these topics. Tse *et al.* [1] designed for use in vibration-based machine fault diagnosis. Fong and Hui [2] described an intelligent data mining technique that combines neural network and rule-based reasoning with case-based reasoning to mine information from the customer service database for online machine fault diagnosis. Liu and Liu [3] presented an efficient expert system for machine fault diagnosis to improve the efficiency of the diagnostic process. Zeng and Wang [4] investigated the feasibility of employing fuzzy sets theory in an integrated machine-fault diagnostic system. Son *et al.* [5] presented a novel research using smart sensor systems for machine fault diagnosis. Kurek and Osowski [6] presented an automatic computerized system for the diagnosis of the rotor bars of the induction electrical motor by applying the support vector machine. Based on [10-12], in this study, we propose two propositions to treat the machine diagnosis fault.

Section 2 is Preliminaries. The proposed diagnosing machine fault method is in Section 3. The example implementation is in Section 4. Finally, we make the conclusion in Section 5.

J.-S. Pan, S.-M. Chen, N.T. Nguyen (Eds.): ACIIDS 2012, Part I, LNAI 7196, pp. 264–269, 2012.

2 Preliminaries

For the proposed algorithm, all pertinent definitions of fuzzy sets are given below [7-12].

Definition 2.1. Triangular Fuzzy Numbers: Let $\tilde{A} = (p,\ q,\ r)$, $p<q<r$, be a fuzzy set on R. It is called a triangular fuzzy number, if its membership function is

$$\mu_{\tilde{A}}(x) = \begin{cases} \dfrac{x-p}{q-p}, & if\ p \le x \le q \\ \dfrac{r-x}{r-q}, & if\ q \le x \le r \\ 0, & otherwise \end{cases} \tag{1}$$

If $r=q$, $p=q$, then $\tilde{A}$ is $(q,\ q,\ q)$. We call it the fuzzy point $\tilde{q}$ at q.

Proposition 2.1. Let $\tilde{A}_1 = (p_1,\ q_1,\ r_1)$ and $\tilde{A}_2 = (p_2,\ q_2,\ r_2)$ be two triangular fuzzy numbers, and k>0, then, we have

$$(1^0)\ \tilde{A}_1 \oplus \tilde{A}_2 = (p_1 + p_2,\quad q_1 + q_2,\quad r_1 + r_2) \tag{2}$$

$$(2^0)\ k \otimes \tilde{A}_1 = (kp_1, kq_1, kr_1) \tag{3}$$

We can easily show the Proposition 2.1 by extension principle.

Definition 2.2. Fuzzy relation: Let $X, Y \subseteq R$ be universal sets, then

$$\tilde{R} = \{((x,y), \mu_{\tilde{R}}(x,y)) | (x,y) \subseteq X \times Y\} \tag{4}$$

is called a fuzzy relation on $X \times Y$.

3 The Proposed Method

The observation of machines working in the normal state or at faulty conditions allows us to point out some typical symptoms indicating the fault causes. These symptoms let us create a diagnostic model of the machine, responsible for early estimation of the technical fault causes of the machine.

Let $S = \{S_1, S_2, \ldots, S_P\}$ be the set of one machine with fault symptoms, and $F = \{F_1, F_2, \ldots, F_q\}$ be the set of fault causes. Also, we let $M = \{M_1, M_2, \ldots, M_r\}$ be the crisp universal set of machines.

Step 1: Let $\tilde{H}$ be the fuzzy relation on the set *M* of machines and the set *S* of symptoms as in Eq. (5).

$$\tilde{H} = \begin{bmatrix} a_{11} & a_{12} \cdots a_{1p} \\ a_{21} & a_{22} \cdots a_{2p} \\ \vdots & \vdots \quad \vdots \\ a_{r1} & a_{r2} \cdots a_{rp} \end{bmatrix} \tag{5}$$

Step 2: Let $\tilde{R}$ be the fuzzy relation on the set *S* of machine with fault symptoms and the set *F* of fault causes as in Eq. (6).

$$\tilde{\Re} = \begin{pmatrix} r_{11} & r_{12} & \cdots r_{1q} \\ r_{21} & r_{22} & \cdots r_{2q} \\ \vdots & \vdots & \vdots \\ r_{\mathrm{p}1} & r_{\mathrm{p}2} & \cdots r_{pq} \end{pmatrix} \tag{6}$$

Step 3: By Eq. (5) and Eq. (6), the fuzzy relation $\tilde{B}$ on the set *M* of machines and the set *F* of fault causes can be inferred by means of the compositional rule of inference as in Eq. (7).

$$\tilde{B} = \tilde{H} \circ \tilde{\Re} = \begin{pmatrix} b_{11} & b_{12} \cdots b_{1\mathrm{q}} \\ b_{21} & b_{22} \cdots b_{2\mathrm{q}} \\ \vdots & \vdots \quad \vdots \\ b_{\mathrm{r}1} & b_{\mathrm{r}2} \cdots b_{\mathrm{rq}} \end{pmatrix} \tag{7}$$

where

$$b_{tv} = \min\left((\sum_{k=1}^{p} a_{tk} \cdot r_{kv}), 1 \right) \tag{8}$$

Normalize b_{tv} in Eq. (8), we let

$$B_{tv} = \frac{b_{tv}}{\sum_{k=1}^{q} b_{tk}} \tag{9}$$

Then, we have

$$0 < B_{tv} \leq 1 \text{ and } \sum_{v=1}^{q} B_{tv} = 1 \tag{10}$$

for *t*=1, 2, …, *r*.

Step 4: From Eq. (9), we have the fuzzy inferred diagnosis of the machine M_t as follows:

$$\tilde{B}_t = \frac{B_{t1}}{F_1} + \frac{B_{t2}}{F_2} + \cdots + \frac{B_{tq}}{F_q} \tag{11}$$

Eq. (11) shows that B_{tv} is the membership grade of the fault cause F_j for the machine M_t, i.e., B_{tv} is the fault grade of the F_j for the machine M_t.

Then, we have the following Proposition 3.1:

Proposition 3.1. By the fuzzy compositional rule of inference, we have

(1^0) The optimal diagnosis based on the maximal membership grade:

We let

$$B_{ts(t)} = \max_{1 \leq v \leq q} B_{tv}, \quad s(t) \in \{1, 2, ..., q\} \tag{12}$$

then the machine M_t is diagnosed fault cause $F_{s(t)}$, for *t*=1, 2, …, *r*.

(2^0) By the probability distribution principle:

For the machine M_t, (*t*=1, 2, …, *r*), the probability of the fault cause F_v is B_{tv}, for *v*=1, 2, …, *q*.

Now, we consider the new machine M_*, ($M_* \notin M$). Let the fuzzy set $\tilde{A} = (a_1, a_2, ..., a_p)$, $0 \leq a_j \leq 1$, *j*=1, 2, …, *p*, be the fuzzy relation on M_* and the set $S = \{S_1, S_2, \ldots, S_P\}$ of fault symptoms. $\tilde{A} = (a_1, a_2, ..., a_p)$ is given by the evaluator/decision maker. Replaced $\tilde{H}$ in Eq. (7) by $\tilde{A}$, we let

$$\tilde{C} = (C_1, C_2, ..., C_q) = \tilde{A} \circ \tilde{R} \tag{13}$$

where

$$C_k = Min(\sum_{t=1}^{p} a_t r_{tk}, 1) \tag{14}$$

Normalize C_k in Eq. (14), we let

$$D_k = \frac{C_k}{\sum_{t=1}^{q} C_t} \in [0,1] \tag{15}$$

for k=1, 2, …, q. Then, we have the following Proposition 3.2:

Proposition 3.2. For the new machine M_*, let $\tilde{A} = (a_1, a_2, ..., a_p)$, $0 \le a_j \le 1$, j=1, 2, …, p, be the fuzzy relation on machine M_* and the set S of fault symptoms. Then, we have

(1^0) By the maximal membership grade:
We let

$$D_m = \max_{1 \le k \le q} D_k \,, \;\; m \in \{1, 2, ..., q\} \tag{16}$$

then the machine M_* is diagnosed fault cause F_m.

(2^0) By the probability distribution principle:
For the machine M_*, the probability of the fault cause F_k is D_k, for k=1, 2, …, q.

4 Numeric Example

If we have $S = \{S_1, S_2, S_3, S_4\}$ the set of one machine with fault symptoms, and $F = \{F_1, F_2, F_3\}$ be the set of fault causes as the following data in Table 1:

Table 1. The data of S_i and F_j

	F_1	F_2	F_3
S_1	0.4	0.35	0.25
S_2	0.41	0.27	0.32
S_3	0.35	0.4	0.25
S_4	0.36	0.29	0.35

If $\tilde{A} = (0.3, 0.4, 0.1, 0.2)$ is given by the decision maker, then by Proposition 3.2, we have D_1=0.391, D_2=0.311, D_3=0.298.

(1^0) By the maximal membership grade, the machine M_* is diagnosed fault cause F_1.

(2^0) By the probability distribution principle, the probability of the fault cause F_1 is 0.391, F_2 is 0.311 and F_3 is 0.298.

5 Conclusion

In this study, we present a fuzzy diagnosing machine fault model. We use the fuzzy compositional rule inference to present two propositions to treat the machine diagnosis fault.

References

1. Tse, P.W., Yang, W.-X., Tam, H.Y.: Machine fault diagnosis through an effective exact wavelet analysis. Journal of Sound and Vibration 277(4-5), 1005–1024 (2004)
2. Fong, A.C.M., Hui, S.C.: An intelligent online machine fault diagnosis system. Journal of Computing & Control Engineering 12(5), 217–223 (2001)
3. Liu, S.C., Liu, S.Y.: An Efficient Expert System for Machine Fault Diagnosis. The International Journal of Advanced Manufacturing Technology 21(9), 691–698 (2003)
4. Zeng, L., Wang, H.P.: Machine-fault classification: A fuzzy-set approach. The International Journal of Advanced Manufacturing Technology 6(1), 83–93 (1991)
5. Son, J.-D., Niu, G., Yang, B.-S., Hwang, D.-H., Kang, D.-S.: Development of smart sensors system for machine fault diagnosis. Expert Systems with Applications 36(9), 11981–11991 (2009)
6. Kurek, J., Osowski, S.: Support vector machine for fault diagnosis of the broken rotor bars of squirrel-cage induction motor. Neural Comput. & Applic. 19, 557–564 (2010)
7. Kaufmann, A., Gupta, M.M.: Introduction to Fuzzy Arithmetic Theory and Application, Van Nortrand, New York (1991)
8. Zimmermann, H.T.: Fuzzy Sets Theory and Its Application. Kluwer Academic Publishers, Boston (1991)
9. Zadeh, L.A.: Fuzzy Sets. Information and Control 8, 338–353 (1965)
10. Lee, H.-M.: Applying Fuzzy Set Theory to Evaluate the Rate of Aggregative Risk in Software Development. Fuzzy Sets and Systems 79, 323–336 (1996)
11. Lin, L., Lee, H.-M.: Fuzzy Assessment Method on Sampling Survey Analysis. Expert Systems with Applications 36, 5955–5961 (2009)
12. Lin, L., Lee, H.-M.: Group Assessment Based on the Linear Fuzzy Linguistics. International Journal of Innovative Computing Information and Control 6(1), 263–274 (2010)

Evaluation of the Improved Penalty Avoiding Rational Policy Making Algorithm in Real World Environment

Kazuteru Miyazaki[1], Masaki Itou[2], and Hiroaki Kobayashi[3]

[1] National Institution for Academic Degrees and University Evaluation
teru@niad.ac.jp
[2] Toshiba Tec Corporation
[3] Meiji University
kobayasi@isc.meiji.ac.jp

Abstract. We focus on a potential capability of Exploitation-oriented Learning (XoL) in non-Markov multi-agent environments. XoL has some degree of rationality in non-Markov environments and is also confirmed the effectiveness by computer simulations. Penalty Avoiding Rational Policy Making algorithm (PARP) that is one of XoL methods was planed to learn a penalty avoiding policy. PARP is improved to save memories and to cope with uncertainties, that is called Improved PARP. Though the effectiveness of Improved PARP has been confirmed on computer simulations, there is no result in real world environment. In this paper, we show the effectiveness of Improved PARP in real world environment using a keepaway task that is a testbed of multi-agent soccer environment.

Keywords: Reinforcement Learning, Exploitaion-oritented Learning, Keepaway Task, Soccer Robot.

1 Introduction

Reinforcement learning (RL)[14] is a kind of machine learning[6,4]. It aims to adapt an agent to a given environment with a reward and a penalty. Traditional RL systems are mainly based on the Dynamic Programming (DP). They can get an optimum policy that maximizes an expected discounted reward in Markov Decision Processes (MDPs). We know Temporal Difference learning [14] and Q-learning [14] as a kind of the DP-based RL systems. They are very attractive since they are able to guarantee the optimality in MDPs. We know that Partially Observable Markov Decision Processes (POMDPs) classes are wider than MDPs. If we apply the DP-based RL systems to POMDPs, we will face some limitation. Hence, a heuristic eligibility trace is often used to treat a POMDP. We know TD(λ) [14] Sarsa(λ) [14] and Actor-Critic [3] as such kinds of RL systems.

The DP-based RL system aims to optimize its behavior under given reward and penalty values. However, it is difficult to design these values appropriately for the purpose of us. If we set inappropiate values, the agent may learn unexpected behavior [9]. We know the Inverse Reinforcement Learning (IRL) [12,1] as a method related to the design problem of reward and penalty values.

J.-S. Pan, S.-M. Chen, N.T. Nguyen (Eds.): ACIIDS 2012, Part I, LNAI 7196, pp. 270–280, 2012.

On the other hand, we are interested in the approach where a reward and a penalty are treated independently. As examples of RL systems in the environment where the number of types of a reward is one, we know the rationality theorem of Profit Sharing (PS) [7], the Rational Policy Making algorithm (RPM) [8] and so on. Furthermore, we know the Penalty Avoiding Rational Policy Making algorithm (PARP) [9] and Improved PARP [15] as examples of RL systems that are able to treat a penalty, too. We call these systems Expolitation-oriented Learning (XoL) [11].

XoL have several features: (1) Though traditional RL systems require appropriate reward and penalty values, XoL only requires the degree of importance among them. In general, it is easier than designing their values. (2) They can learn more quickly since they trace successful experiences very strongly. (3) They are not suitable for pursuing an optimum policy. The optimum policy can be acquired with multi-start method [8] but it needs to reset all memories to get a better policy. (4) They are effective on the classes beyond MDPs since they are a Bellman-free method [14] that do not depend on DP.

We are interested in XoL since we require quick learning and/or learning in the class wider than MDPs. Especially, we focus on Improved PARP whose effectiveness has been confirmed on computer simulations [15,5]. However there is no result in real world environment. In this paper, we show the effectiveness of Improved PARP in real world environment using Keepaway task [13] that is a testbed of multiagent soccer environment.

2 The Domain

2.1 Notations

Consider an agent in some unknown environment. For each discrete time step, after the agent senses the environment as a pair of a discrete attribute and its value, it selects an action from some discrete actions and executes it. In usual DP-based RL systems and PS, a scalar weight, that indicates the importance of a rule, is assigned to each rule. The environment provides a reward or a penalty to the agent as a result of some sequence of actions.

We term the sequence of rules selected between the rewards as an episode. Consider a part of an episode where the sensory input of the first selection rule and the sensory output of the last selection rule are the same although both rules are different. We term it a detour. The rules on a detour may not contribute to obtain a reward. We term a rule irrational if and only if it always exists on detours in any episodes. Otherwise, the rule is termed rational. We term a rule a penalty rule if and only if it gets a penalty or the state transit to a penalty state in which there are penalty or irrational rules only.

The function that maps sensory inputs to actions is termed a policy. The policy that maximizes the expected reward per an action is termed as an optimum policy. We term a policy rational if and only if the expected reward per an action is positive.We term a rational policy a penalty avoiding rational policy if and only if it has no penalty rule. Furthermore, the policy that outputs just one action for each sensory input is termed a deterministic policy.

2.2 The Penalty Avoiding Rational Policy Making Algorithm

We know the Penalty Avoiding Rational Policy Making algorithm (PARP) as XoL that can make a penalty avoiding rational policy. To avoid all penalties, PARP suppresses all penalty rules in the current rule sets with the Penalty Rule Judgment algorithm (PRJ) in Fig. 2. After suppressing all penalty rules, it aims to make a deterministic rational policy by PS, RPM and so on.

```
procedure The Penalty Rule Judgement (PRJ)
begin
  Put the mark in the rule that has been got a penalty directly
  do
    Put the mark on the following state ;
      there is no rational rule   or   there is no rule that can
      move the state to any non-marked states.
    Put the mark in the following rule ;
      there are marks in the states that can be transited by it.
  while (there is a new mark on some state)
end.
```

Fig. 1. The Penalty Rule Judgment algorithm (PRJ); We can regard the marked rule as a penalty rule. We can find all penalty rules in the current rule set through continuing PRJ.

Furthermore, PARP avoids a penalty stochastically if there is no deterministic rational policy. It is realized by selecting the rule whose penalty level is the least in all penalty levels at the same sensory input. The penalty level is estimated by the transition probability to a penalty state of each rule. If we can continue to select only rational rules, we will be able to get a penalty avoiding rational policy. On the other hand. if we have to refer to the penalty level, we will get a policy that has a possibility to get a penalty.

PARP uses PRJ but it has to memorize all rules that have been experienced and descendant states that have been transited by their rules to find a penalty rule. It requires O(MN^2) memories where N and M are the number of types of a sensory input and an action. Furthermore, PARP requires the same order of memories to suppress all penalties stochastically. In applying PRJ to large-scale problems, we are confronted with the curse of dimensionality.

2.3 Improved PARP

Improved PARP can find a penalty rule by $O(MN)$ memories where PRJ is applied only to the episode that ends with penalty.

Furthermore, *penalty rule rate* of a rule is introduced to estimate the transition probability to a penalty state when the rule is applied. This parameter is used to cope with the uncertainty of the state transition. Penalty rule rate $PL(xa)(0 \leq PL(xa) \leq 1.0)$ of rule "xa" is given by the following equation,

$$PL(xa) = \frac{N_p(xa)}{N(xa)}, \tag{1}$$

where $N_p(xa)$ is the number of times which was judged as a penalty rule by PRJ, and $N(xa)$ is the number of times which was chosen untill then. If $PL(xa)$ is getting closer to 1, the rule "xa" has higher possibility of receiving penalty. Conversely, if $PL(xa)$ closes to 0, the selected action will seldom receive a penalty.

Now, let γ be *threshold* of the penalty rule rate, then rules are classified as follows:

$$rule\ xa = \begin{cases} penalty\ rule & (\text{if } PL(xa) > \gamma) \\ non\ penalty\ rule & (\text{otherwise}) \end{cases} \tag{2}$$

This means that large γ will reduce the number of penalty rules and increase the number of applicable rules in the state.

3 Keepaway Task

3.1 Basic Setting

In the Robocup soccer Keepaway environment [13] studied as a multiagent consecutive task benchmark,a keeper agent tries to keep a ball as long as possible without being stolen by the opponent agent called a taker, requiring cooperative behavior among keeper agents.

Though there are several researches [2,15] that use the keepaway simulator provided by the paper [13], we execute an experiment in real world environment using small robots. We use 3 keepers and a taker. This 3-vs-1 keepaway task occupies a 230 [cm] × 230 [cm] playing field (Fig. 2). The ball-holding keeper is K_1 and the other keepers K_2 and K_3 based on whichever is closer to K_1. The three keepers are initially located as shown in Fig. 2 where positions of K_1, K_2 and K_3 are (0,-150), (-100,-150) and (100,150). The taker T_1 and the ball are located at (0,150) and (0,-100) initially. Only keepers are learning agents and learn their policies.

3.2 State Variables

The state variables are composed of angular relationships between K_1 and K_2, between K_1 and K_3, between K_1 and the ball, and between K_1 and T_1 as shown in Fig.3. Furthermore, six distance labels among K_1 and the other keepers are added to the state variables as shown in Table. 1 where label "1" means that K_2 is the closest to K_1, and T_1 is the farthest. We have $4^4 \times 6 = 1536$ states. Only the keeper that holds the ball receives state variables.

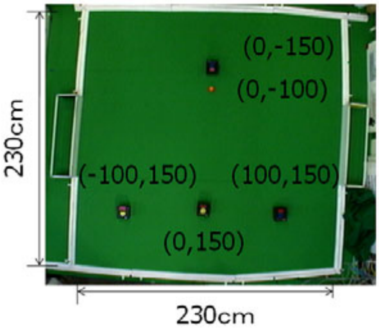


Fig. 2. Initial positions

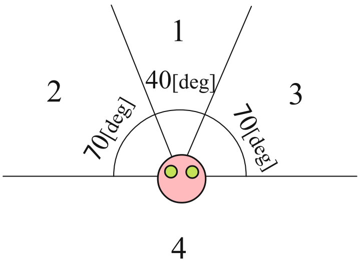


Fig. 3. Direction labels

Table 1. Distance labels

	label					
	1	2	3	4	5	6
K_2	1	1	2	3	2	3
K_3	2	3	1	1	3	2
T_1	3	2	3	2	1	1

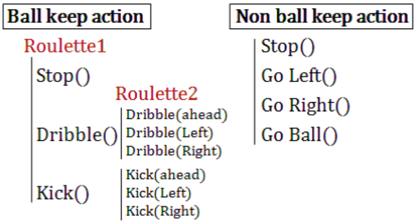


Fig. 4. Agent action roulette

3.3 Actions

Following five macro actions are defined:

Stop(): stay at the current location.
Dribble(α): dribble in the direction α.
Kick(α): kick in the direction α.
Go Ball(): turn to the ball and moves one step.
Go Left(): turn by 45 degrees to the left and move one step.
Go Right(): turn by 45 degrees to the right and move one step.

α is selected from the set $\{ahead, left, right\}$ where "ahead" means the front direction and "left" or "right" is 45 degrees to the left or right. Dribble() and Kick() are achieved by pushing the ball forward, but the pushing strenght of Dribble() is one third of Kick().

K_1, the ball keeper, selects the action by two stages as shown in Fig. 4. K_1 selects an action from the set { Stop(), Dribble(), Kick() }. It is decided with Roullet1 where the maximum weights of { Dribble(ahead), Dribble(Left), Dribble(Right) } and { Kick(ahead), Kick(Left), Kick(Right) } are assigned to the weights of Dribble() and Kick(), respectively. If the selected action is one of Dribble() or Kick(), the detail action is decided with another roulette (Roullet2).

Table 2. Combinations of Colors for Robots

K_1	blue+red
K_2	yellow+pink
K_3	blue+pink
T_1	yellow+red

Table 3. RGB Parameter of Each Color

Color	R	G	B
orange	110 – 255	50 – 140	0 – 65
blue	0 – 40	0 – 70	45 – 200
yellow	150 – 255	130 – 255	20 – 170
red	110 – 255	50 – 140	0 – 64
pink	95 – 255	45 – 120	60 – 150

K_2 and K_3 selects an action from the set { Stop(), Go Ball(), Go Left(), GoRight() } randomly. T_1 always uses "Go Ball()." Each agent selects an action in the order of K_1, K_2, K_3 and T_1. If one action has been completed, the next agent selects and executes an action.

4 The Soccer Robot System

We have developed a soccer robot system (Fig.5) to evaluate impoved PARP. It has the decision making subsystem and the image processing subsystem. The former is to decide an action and the latter is to process images from camera. It can adapt on the reguration of Robo Cup Small Size Robot League except for the field size.

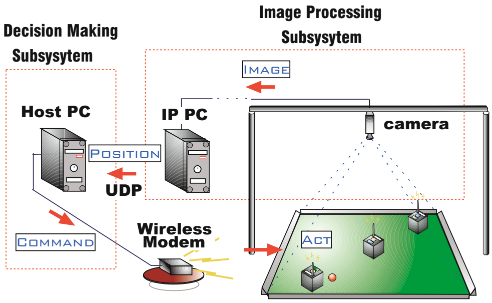


Fig. 5. Soccer Robot System

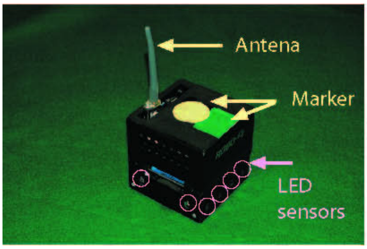


Fig. 6. Soccer Robot Robo-E

4.1 Soccer Robot

We use three robots called Robo-E (Fig.6). It has 12 LED sensors. They are used to escape from clashing with a wall or other robots.

Each robot has different markers each other to distinguish each robot. We show the combinations of colors for robots in Table.2.

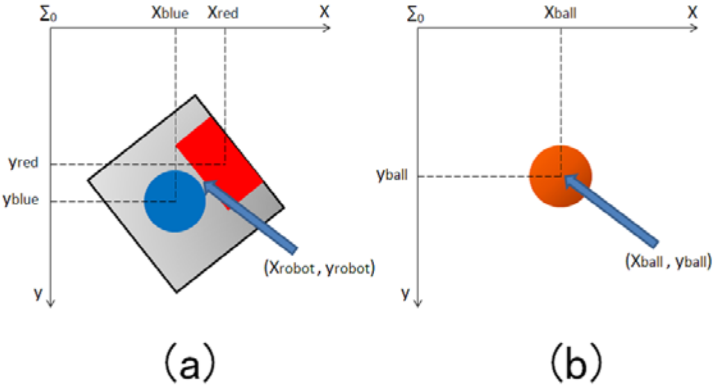


Fig. 7. Positions of Robots and a Ball

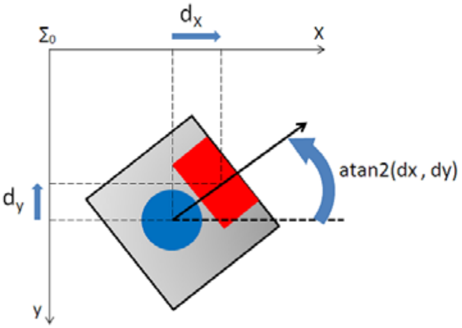


Fig. 8. Orientation of a Robot

4.2 Host System

The host system is consturcted by two computers, that is, Host and Image Processing (IP) PCs. IP and Host PCs are connected by TCP/IP each other. We use UDP (User Datagram Protocol) as the protocol on IP.

Host PC is to decide an action. We have impremented Improved PARP on Host PC. On the other hand, IP PC calculates positions of robots and a ball. Their images are given by the camera (SONY XC-003). The camera is mounted at the top of the field to get global vision. We use machine vision tool HALCON with its Hdevelop programming environment to get robots and a ball positions.

4.3 Position and Orientation

The positions of each robots are calculated with markers (blue, yellow, red and pink) and their body color (black). We use RGB parameter in the Table 3 to distinguish each color.

The position of the robot is at the center of gravity of two markers : (Xrobot, Yrobot) in Fig.7(a). The position of a ball is at the center of ball marker (orange): (Xball, Yball) in Fig.7(b).

The orientation of the robot is calculated by the center of gravity of two markers (Fig.8). Let the coordinate of the center of gravity of center marker be (Xcenter, Ycenter) and another one be (Xoutside, Youtsde), we can get the angle θ by $\theta = \text{atan2}(d_y, d_x)$ $(-180° < \theta < 180°)$ where $d_x = x_{outside} - x_{center}$ and $d_y = y_{outside} - y_{center}$.

5 Real World Experimentation

5.1 Setting

We perform the experiments according to the method described in previous sections. The sampling time is 150[msec]. One experiment is continued unitl 150 trials where a robot selects an action 150 times. We execute 5 experiments with different random seeds.

5.2 Reward and Penalty

The initial weight value of each rule is set to 10.0. If the pass from K_1 to either K_2 or K_3 is successful, reward 100.0 is given. The reward is shared among rules in the episode with discount rate 0.8.

The experiment ends, if Taker T_1 steals the ball or the ball goes out of the playing field. Dribble(α) and Kick(α) require to move behind the ball to kick the ball. If agent K_1 cannot move there, the experiment also ends. In these cases, penalty is given to the last rule and PRJ algorithm is activated. All $N_p(xa)$s in Eq.1 of marked rule "xa"s are increased by one. We set $\gamma = 0.6$.

5.3 Results

Fig.9 shows the number of successful passes every 10 trials for 5 experiments plotted againt the number of trials. Fig.10 is the average number of successful passes of the 5 experiments of Fig.9.

In the case of ①, K_1 learned "Kick(right)," that is the best action as described later, within 30 trials. In ②, the agent learned "Kick(right)" and "Dribble(right)" at the same time; this resulted in a little inferior performance. In ③, K_1 learned "Kick(left)" at first and the agent learned "Kick(right)" after "Kick(left)" was classified as a penalty rule. In ④, K_1 could not learn "Kick(right)" within 150 trials since the agent failed to get enough reward. In ⑤, K_1 could learn "Kick(right)" within 30 trials, but "Kick(right)" was classified as a penalty rule since the action often failed due to the uncertainty.

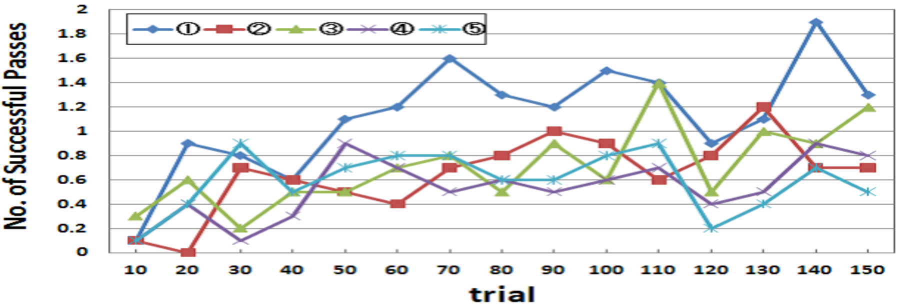


Fig. 9. The number of successful passes each 5 experiment

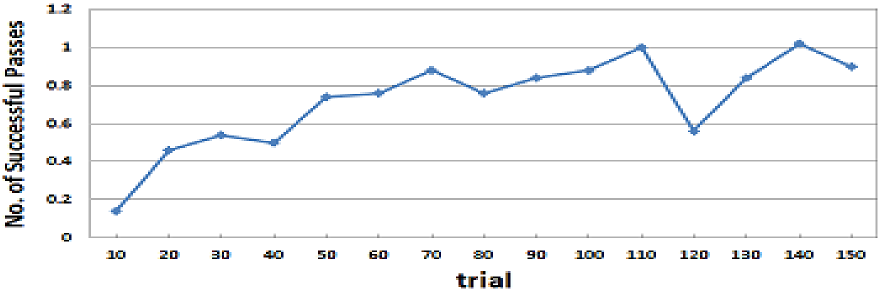


Fig. 10. The average number of successful passes

5.4 Discussion

We can recognize the effect of learning with a penalty from Fig.10. Namely, since "Kick(ahead)" is more likely to fail, it was regarded as a penalty rule in the early trials and the rule was removed from the candidates of the action to accelerate the learning speed.

In all experimants, "Kick(right)", "Kick(left)" or "Dribble(ahead)" were learned with almost same probability in the initial state. "Kick(right)" is better than "Kick(left)" since there is a slope such that "Kick(right)" is preferred in our environment. On the other hand, "Dribble(ahead)" is helpful to avoid the taker. Therefore, to learn "Kick(right)" in the early trials is very profitable for the agents.

① is the best result. ② and ③ are reasonable results, too. But ④ and ⑤ failed to learn in 150 trials. ④ was occured since it was difficult for the agent to get enough reward in 150 trials. On the other hand, if we use larger γ, we would be able to avoid the case of ⑤

PARP required more trials than Improved PARP in our previous computer simulations [15,5]. It could not improve the number of successful passes within 150 trials on this task, too.

6 Conclusion

We focused on a potential capability of Exploitation-oriented Learning (XoL) in non-Markov multi-agent environments. Penalty Avoiding Rational Policy Making algorithm (PARP) that is one of XoL methods was planed to learn a penalty avoiding policy. PARP is improved to save memories and to cope with uncertainties, and is called Improved PARP. In this paper, we have shown the effectiveness of Improved PARP in real world environment using a keepaway task that is a testbed of multiagent soccer environment.

In the future, we will treat continuous sensory inputs and actions directly [10]. Furthermore, we will apply our method to our real biped robot [5].

References

1. Abbeel, P., Ng, A.Y.: Exploration and apprenticeship learning in reinforcement learning. In: Proc. of the 22nd International Conference on Machine Learning, pp. 1–8 (2005)
2. Arai, S., Tanaka, N.: Experimental Analysis of Reward Design for Continuing Task in Multiagent Domains – RoboCup Soccer Keepaway. Transactions of the Japanese Society for Artificial Intelligence 21(6), 537–546 (2006) (in Japanese)
3. Kimura, H., Kobayashi, S.: An analysis of actor/critic algorithm using eligibility traces: reinforcement learning with imperfect value function. In: Proc. of the 15th Int. Conf. on Machine Learning, pp. 278–286 (1998)
4. Hong, T., Wu, C.: An Improved Weighted Clustering Algorithm for Determination of Application Nodes in Heterogeneous Sensor Networks. J. of Information Hiding and Multimedia Signal Processing. 2(2), 173–184 (2011)
5. Kuroda, S., Miyazaki, K., Kobayashi, H.: Introduction of Fixed Mode States into Online Profit Sharing and Its Application to Waist Trajectory Generation of Biped Robot. In: European Workshop on Reinforcement Learning 9 (2011)
6. Lin, T.C., Huang, H.C., Liao, B.Y., Pan, J.S.: An Optimized Approach on Applying Genetic Algorithm to Adaptive Cluster Validity Index. International Journal of Computer Sciences and Engineering Systems 1(4), 253–257 (2007)
7. Miyazaki, K., Yamamura, M., Kobayashi, S.: On the Rationality of Profit Sharing in Reinforcement Learning. In: Proc. of the 3rd Int. Conf. on Fuzzy Logic, Neural Nets and Soft Computing, pp. 285–288 (1994)
8. Miyazaki, K., Kobayashi, S.: Learning Deterministic Policies in Partially Observable Markov Decision Processes. In: Proc. of 5th Int. Conf. on Intelligent Autonomous System, pp. 250–257 (1998)
9. Miyazaki, K., Kobayashi, S.: Reinforcement Learning for Penalty Avoiding Policy Making. In: Proc. of the 2000 IEEE Int. Conf. on Systems, Man and Cybernetics, pp. 206–211 (2000)
10. Miyazaki, K., Kobayashi, S.: A Reinforcement Learning System for Penalty Avoiding in Continuous State Spaces. J. of Advanced Computational Intelligence and Intelligent Informatics 11(6), 668–676 (2007)
11. Miyazaki, K., Kobayashi, S.: Exploitation-Oriented Learning PS-r$^{\#}$. J. of Advanced Computational Intelligence and Intelligent Informatics 13(6), 624–630 (2009)
12. Ng, A.Y.,, Russell, S.J.: Algorithms for Inverse Reinforcement Learning. In: Proc. of the 17th Int. Conf. on Machine Learning, pp. 663–670 (2000)

13. Stone, P., Sutton, R.S., Kuhlamann, G.: Reinforcement Learning toward RoboCup Soccer Keepaway. Adaptive Behavior 13(3), 0165–0188 (2005)
14. Sutton, R.S., Barto, A.G.: Reinforcement Learning: An Introduction. A Bradford Book. MIT Press (1998)
15. Watanabe, T., Miyazaki, K., Kobayashi, H.: A New Improved Penalty Avoiding Rational Policy Making Algorithm for Keepaway with Continuous State Spaces. J. of Advanced Computational Intelligence and Intelligent Informatics. 13(6), 675–682 (2009)

Similarity Search in Streaming Time Series Based on MP_C Dimensionality Reduction Method

Thanh-Son Nguyen[1] and Tuan-Anh Duong[2]

[1] Faculty of Information Technology, HCM City University of Technical Education
[2] Faculty of Computer Science and Engineering, HCM City University of Technology

Abstract. The similarity search problem in streaming time series has become a hot research topic since such data arise in so many applications of various areas. In this problem, the fact that data streams are updated continuously as new data arrive in real time is a challenge due to expensive dimensionality reduction recomputation and index update costs. In this paper, adopting the same ideas of a delayed update policy and an incremental computation from IDC index (Incremental Discrete Fourier Transform (DFT) Computation – Index) we propose a new approach for similarity search in streaming time series by using MP_C as dimensionality reduction method with the support of Skyline index. Our experiments show that our proposed approach for similarity search in streaming time series is more efficient than the IDC-Index in terms of pruning power, normalized CPU cost and recomputation and update time.

1 Introduction

A streaming time series (STS) is a real-value sequence $C = c_1, c_2, \ldots,$ where new values are continuously added at the end of the sequence C as time progresses. Because a STS includes a great number of values, similarity is measured with the W last values of the streams (W is the length of a sliding window).

The similarity search problem in STS has become a hot research topic due to its importance in many applications of various areas such as earthquake forecast, internet traffic examination, moving object examination, financial market analysis, and anomaly detection ([2],[5],[6]). The challenge in these applications is that STS comes continuously in real time, i.e., time series data are frequently updated. Methods that are used to perform similarity search on archived time series in the past may not work efficiently in streaming scenarios because the update and recomputation costs in STS are significant. Therefore one needs an efficient and effective method for similarity search in this time series data type.

In [5] and [6], Kontaki et al. proposed an index structure, IDC-Index (Incremental DFT Computation – Index), which can be used for the similarity search in STS. The IDC-Index is based on a multi-dimensional index structure, R*-tree, improved with a delayed update policy and an incremental calculation of DFT. This approach is used in order to reduce update and re-computation costs in STS similarity search. However, the effect of using IDC-Index is not high due to using R*-tree as index structure and DFT as dimensionality reduction method. In order to enhance the efficiency of similarity search

J.-S. Pan, S.-M. Chen, N.T. Nguyen (Eds.): ACIIDS 2012, Part I, LNAI 7196, pp. 281–290, 2012.

in STS we propose a new approach which uses MP_C, as dimensional reduction method and Skyline index as multidimensional index structure.

In the proposed approach, while using the same idea of a delayed update strategy and an incremental calculation for feature extraction, we can show that our new method for time series dimensionality reduction, MP_C, with the support of Skyline index can provide a more efficient similarity search in STS than IDC-Index in terms of pruning power, normalized CPU cost and recomputation and update time.

2 Preliminaries

2.1 Index Structures

The popular multidimensional index structures are R-tree and its variants ([1], [3]). In a multidimensional index structure (e.g., R-tree or R^*-tree), each node is associated with a minimum bounding rectangle (MBR). A MBR at a node is the minimum bounding box of the MBRs of its child nodes. A potential weakness in the method using MBR is that MBRs in index nodes can overlap. Overlapping rectangles could have negative effect on the search performance. Besides, another problem in the method using MBR is that summarizing data in MBRs, the sequence nature of time series is not captured.

Skyline Index, another elegant paradigm for indexing time series data which uses another kind of minimum bounding regions, is proposed by Li et al., 2004 [9]. Skyline Index adopts new Skyline Bounding Regions (SBR) to approximate and represent a group of time series data according to their collective shape. An SBR is defined in the same *time-value* space where time series data are defined. Therefore, SBRs can capture the sequential nature of time series. SBRs allow us to define a distance function that tightly lower-bounds the distance between a query and a group of time series data. SBRs are free of internal overlaps. Hence using the same amount of space in an index node, SBR defined a better bounding region. For k-nearest-neighbor (KNN) queries, Skyline index approach can be coupled with some well-known dimensionality reduction technique such as APCA and improve its performance by up to a factor of 3 ([9]).

2.2 Similarity Search in Streaming Time Series

Many solution approaches have been proposed for similarity query in streaming time series lately. In [4] Gao and Wang, 2002, proposed a method based on a prediction for similarity search in STS. In this case, time series data are static and the query is changes over time. The authors solve the problem by using Fast Fourier Transform (FFT) to find out the cross correlations between the query and time series. The Euclidean distance between the query and each time series is calculated based on the predicted values. When the actual query is incoming the prediction error and the predicted distances are used to discard false alarms.

In [8], Liu et al., 2003, proposed a model for processing STS based on an index structure that can adapt to the change in the length of data objects. In this work, the VA-stream and VA^+-stream index structures are used to query k-nearest neighbors. These methods partition the data space into 2^b cells, where b is defined by user. They distribute different number of bits for each dimension so that the total of these bits is equal to *b*. Each cell contains a bit sequence of length *b* which is an approximation of the data points falling into this cell. When a new value arrives a bit reallocation is performed in order to adjust the structure.

To perform similarity search over high speed time series streams, Lian et al. ([7]), 2007, proposed an approach which based on a multi-scaled segment mean (MSM) as a dimensionality reduction method. MSM representation can be incrementally computed with low cost and helps to limit false alarms and avoid false dismissals.

In [5] and [6], Kontaki et al. proposed an index structure, IDC-Index (Incremental DFT Computation – Index), which is used for the similarity search problem in STS. The IDC-Index is based on a multi-dimensional index structure, R*-tree, improved with a delayed update policy and an incremental calculation of DFT. These authors solved ε-range and *k*-NN search problems in which both query and time series data are dynamic. In this work, the approach will be used as a baseline to compare with our approach because we also adopt the same idea of a delayed update strategy and an incremental calculation as in IDC-Index approach. In the next section, we will have a brief description of this approach.

2.3 IDC-Index

In this approach, DFT method is used for extracting features of streaming time series and R*-tree, based on MBR is used as a multi-dimensional index structure for efficient similarity search. To overcome the difficulties incurred by the streaming environment, these authors used an incremental calculation of DFT in order to avoid re-calculation every time a new value arrives and a delayed update policy in order to avoid updating R*-tree continuously.

The incremental calculation of DFT is compute by the following formula ([6]):

$$\mathrm{DFT}_n(T) = \frac{1}{\sqrt{W}} \cdot \left(\sqrt{W} \cdot \mathrm{DFT}_n(S) - S(0) + T(W)\right) \cdot e^{j2\pi n/W}, \quad (0 \leqslant n \leqslant W-1)$$

where, $S = (S(0), S(1), \ldots, S(W\text{-}1))$ is streaming time series of length *W*, $\mathrm{DFT}_0(S)$, $\mathrm{DFT}_1(S)$, …, $\mathrm{DFT}_{W\text{-}1}(S)$ are Fourier coefficients of *S*, $T= (T(1), T(2), \ldots, T(W))$ is a new sequence when a new value arrives, where $S(i) = T(i)$, for $1 \le i \le W$-1 and $T(W)$ is a new value.

The delayed update strategy is performed as follows. When a new value arrives, a new sequence is created. If the distance between the old DFT vector and the new DFT vector is greater than a given threshold Δ (the threshold Δ is used to direct the updates) then the R*-tree is updated. Otherwise, no update is carried out. Kontaki et al. [5] have already proved that if the query threshold ε is expanded by Δ in the non-leaf nodes of R*-tree then similarity searches lead no false dismissals.

3 MP_C Representation

MP_C – Compression and Clipping

Given a time series *C* and a query *Q*, without loss of generality, we assume *C* and *Q* are *n* units long. *C* is divided into segments. Some points in each segment are chosen. To reduce space consumption, the chosen points are transformed into a sequence of bits, where 1 represents above the segment average and 0 represents below, i.e., if μ is the mean of segment C, then

$$c_t = \begin{cases} 1 & \text{if } c_t > \mu \\ 0 & \text{otherwise} \end{cases}$$

The mean of each segment and the bit sequence are recorded as segment features.

We can choose some points in each segment with different algorithms in time order. For example, in order to choose l points we can extract the first or the last l points of each segment and so on. For the simplicity and the ability of recording the approximate shape of the sequence, in our method, we use the following simple algorithm: (1) dividing each segment into sub-segments, and (2) choosing the middle point of each sub-segment. Fig. 1 shows the intuition behind this technique, with l = 6. In this case, the sequence of bit 010111 and the μ value are recorded.

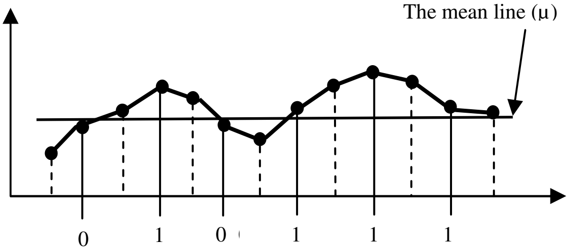


Fig. 1. An illustration of IPIP method

Similarity Measure Defined for MP_C

In order to guarantee no false dismissals we must produce a distance measure in the reduced space, D_{MP_C} which is less than or equal to the distance measure in the original space.

Definition 1(MP_C Similarity Measure). Given a query Q and a time series C (of length n) in raw data. Both C and Q are divided into N segments ($N << n$). Suppose each segment has the length of w. Let C' be an MP_C representation of C. The distance measure between Q and C' in MP_C space, $D_{MP\_C}(Q, C')$, is computed as follows.

$$D_{MP\_C}(Q,C') = \sqrt{D_1(Q,C') + D_2(Q,C')} \quad (1)$$

$D_1(Q, C')$ and $D_2(Q, C')$ are defined as

$$D_1(Q,C') = \sum_{i=1}^{N} w(q\mu_i - c\mu_i)^2 \quad (2)$$

$$D_2(Q,C') = \sum_{j=1}^{N}\sum_{i=1}^{l}(d(q_i,bc_i))^2 \tag{3}$$

where

$q\mu_i$ is the mean value of the i-th segment in Q, $c\mu_i$ is the mean value of the i-th segment in C, bc_i is binary representation of $c_{i.}$ $d(q_i, bc_i)$ is computed by the following formula:

$$d(q_i,bc_i) = \begin{cases} q_i' & \text{if } (q_i' > 0 \text{ and } bc_i = 0) \\ & \text{or} \\ & (q_i' \leq 0 \text{ and } bc_i = 1) \\ 0 & \text{otherwise} \end{cases} \tag{4}$$

q_i' is defined as $q_i' = q_i - q\mu_k$, where q_i belongs to the k^{th} segment in Q.

Lemma 1. If $D(Q, C)$ is the Euclidean distance between query Q and time series C, then $D_{MP\_C}(Q, C') \leq D(Q, C)$.

The proof of Lemma 1 can be seen in our previous paper [10].

4 A Skyline Index for MP_C

In traditional multidimensional index such as R*-tree, minimum bounding rectangles (MBRs) are used to group time series data which are mapped into points in a low dimensional feature-space. If a MBR is defined in the two-dimensional space in which a time series data exists, the overlap between MBRs will be large. So by using the ideas from Skyline index, we can represent more accurately the collective shape of a group of time series data with tighter bounding regions. To attain this aim, we use MP_C bounding regions (MP_C_BRs) for bounding a group of time series data.

Definition 2 (MP_C Bounding Region). Given a group C' consisting of k MP_C sequences in a N-dimensional feature space. The MP_C_BR R of C', is defined as

$R = (C'_{max}, C'_{min})$

where $C'_{max} = \{c'_{1max}, c'_{2max},\ldots,c'_{Nmax}\}$, $C'_{min} = \{c'_{1min}, c'_{2min},\ldots,c'_{Nmin}\}$
and, for $1 \leq i \leq N$, $c'_{imax} = \max\{c'_{i1}, \ldots, c'_{ik}\}$ and $c'_{imin} = \min\{c'_{i1}, \ldots, c'_{ik}\}$
where c'_{ij} is the i^{th} mean value of the j^{th} MP_C sequence in C'.

Figure 2 illustrates an example of MP_C_BR. In this example, BC_i is a bit sequence of time series C_i and the number of middle points in each segment is one.

Once the Skyline Index for MP_C has been built, we have to define the distance function $D_{region}(Q, R)$ of the query Q from the MP_C_BR R associated with a node in the index structure such that it satisfies the group lower-bound condition $D_{region}(Q, R) \leq D(Q, C)$, for any time series C in the MP_C_BR R. The proof of this group lower-bound condition is given in our previous work [10].

We can index the MP_C representation of time series data by first building a Skyline index which based on a spatial index structure such as R*-tree [1]. Each leaf node in the R*-tree-based Skyline index contains an MP_C sequence and a pointer refer to an original time series data in the database. The MP_C_BR associated with a non-leaf node is the smallest bounding region that spatially contains the MP_C_BRs associated with its immediate children.

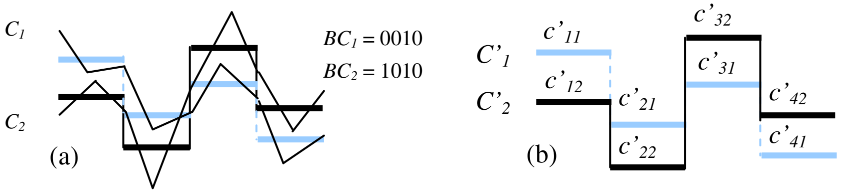


Fig. 2. An example of MP_C_BR. (a)Two time series C_1, C_2 and their approximate MP_C representations in four dimensional space. (b) The MP_C_BR of two MP_C sequences C'_1 and C'_2. $C'_{max} = \{c'_{11}, c'_{21}, c'_{32}, c'_{42}\}$ and $C'_{min} = \{c'_{12}, c'_{22}, c'_{31}, c'_{41}\}$

Two searching problems which we apply in our experiments are ε-range search and KNN search algorithms.

5 Similarity Search in Streaming Time Series Based on MP_C and Skyline Index

In STS new data values arrive continuously and the number of STS in the database may be very large. When a new value arrives, MP_C approximation for this time series must be recomputed using the last W values of this time series. Therefore, the costs in feature extraction recomputation may be high. Besides, the multi-dimensional index structure must be updated every time a new value for a streaming time series arrives. It may lead to a high overhead due to continuous deletions and insertions in the index structure. In order to reduce these costs we apply an incremental computation of MP_C method and a delayed update strategy.

- **The Incremental Computation of MP_C Method**

Let $C = (c_0, c_1, \ldots, c_{n-1})$ is the last sequence of length n of a streaming time series. Suppose C is divided into N segments. The N segment mean values and the middle points of each segment which are transformed into a bit sequence are recorded as features of sequence. Let $C' = (c'_0, c'_1, \ldots, c'_{N-1})$ and a bit sequence bc are the representation of C in MP_C space. When a new value c_n arrives we get a new sequence $S = (s_1, s_2, \ldots, s_n)$, where $c_i = s_i$, $i = 1,.., n$-1 and s_n ($s_n = c_n$) is a new value. The MP_C sequence $S' = (s'_0, s'_1, \ldots, s'_{N-1})$ of S is computed by the previous MP_C approximation of C. This incremental calculation of MP_C method is calculated by the following formula:

$$s'_i = c'_i - \frac{N}{n} c_{\frac{n}{N}i} + \frac{N}{n} c_{\frac{n}{N}(i+1)}$$

And the middle point values, mp_i, of segments are extracted at positions

$$\left\lfloor \frac{n}{N}i + \frac{n}{2N} \right\rfloor$$

where $i = 0, \ldots, N-1$

The chosen middle points are transformed into a sequence of bits, where 1 represents above the average of the corresponding segment and 0 represents below, i.e., if μ_i is the mean of segment C_i, then

$$bc_i = \begin{cases} 1 & \text{if } mp_i > \mu_i \\ 0 & \text{otherwise} \end{cases}$$

It is easy to see that

$$s'_i = c'_i - \frac{N}{n} c_{\frac{n}{N}i} + \frac{N}{n} c_{\frac{n}{N}(i+1)} = \frac{N}{n} \sum_{j=\frac{n}{N}(i-1)+1}^{\frac{n}{N}i} s_j$$

In order to perform the incremental calculation of MP_C method, the last computed MP_C sequences of all streaming time series must be stored.

- **The Delayed Update Strategy in Skyline Index**

According to the delayed update strategy, to prevent a continuous update of STS in Skyline index when a new value arrives, an update threshold T is used to direct the number of updates. To mitigate the cost of update, there is an additional link from time series to a corresponding leaf node in Skyline index. Thank to these "stream to leaf" links, the update can be performed faster from a leaf node to the root.

Suppose C is a STS and C_1= (c_{n-w+1}, … , c_n) is last sequence of length w of C, where c_n is the last value of C. Let C'_1 be a MP_C representation of C_1. When a new value, c_{n+1}, arrives, we get a new sequence C_2= (c_{n-w+2}, … , c_{n+1}) and C'_2, a MP_C representation of C_2, is calculated by the incremental computation based on C'_1. Suppose that MP_C sequence C'_1 is the last sequence which is updated into Skyline index corresponding to C_1. If the distance between the new MP_C representation C'_2 and the old MP_C approximation C'_1 is less than or equal to the update threshold T ($D_{MP\_C}(C'_1, C'_2) \le T$), C'_2 is not updated into Skyline index but it is recorded as the most recent MP_C approximation of the streaming time series C which is used to compute incrementally a new MP_C when a new value arrives.

Let C_3= (c_{n-w+3}, …, c_{n+2}) be a new sequence when another new value, c_{n+2}, arrives and C'_3, a MP_C representation of C_3, is calculated incrementally based on C'_2. C'_3 is recorded as the most recent MP_C approximation instead of C'_2. If the distance $D_{MP\_C}(C'_1, C'_3) > T$, Skyline index is updated by replacing C'_1 with C'_3 in the leaf node and all MP_C_BRs on the path from this leaf node to the root are recomputed. Otherwise, no action is done in the index structure.

In brief, we need both the previously calculated MP_C sequence and the last recorded MP_C sequence. The former is used for the incremental calculation of the new MP_C representation and the latter is used for making a decision whether a MP_C sequence is updated in the index structure or not. Besides, to mitigate further the cost of update, there is an additional link from time series to its corresponding leaf node in Skyline index.

6 Experimental Results

In this section we report the experimental results on similarity search in streaming time series based on MP_C and Skyline index. We compare our proposed technique MP_C using Skyline index (MP_C + Skyline) to the previous approach IDC-Index based on R*-tree. We perform all tests over different reduction ratios and datasets of different lengths. We consider a length of 1024 to be the longest query. Time series datasets for experiments come from various sources publicly available through the Internet and are organized into two separate datasets. The two datasets are Stock-data (37MB - over two million points), Consumer-data (27MB). In order to have an acceptable number of streams, new ones are created by rearranging values of the real streams. The comparison between our proposed approach and IDC-Index is based on the pruning power and the implemented system. Besides, we also compare the index building time, the incremental calculation and the delayed update time of MP_C+ Skyline to that of IDC-Index. In our experiments, the parameters are set: the query distance $\varepsilon = 10$ and the threshold $T = 5$. For brevity, we just show the experimental results on one dataset (Consumer-data).

Pruning Power. In order to compare the effectiveness of two dimensionality reduction techniques, we need to compare their pruning powers. Pruning power P is the fraction of the database that must be examined before we can guarantee that an ε-match to a query has been found. This ratio is based on the number of times we cannot perform similarity search on the transformed data and have to check directly on the original data to find nearest match.

$$P = \frac{\textit{Number of sequences that must be examined}}{\textit{Number of sequences in database}}$$

Figure 3 shows the experimental results of P. In the charts of Figure 3, the horizontal axis represents the value of reduction ratios and the vertical axis represents the pruning power. When the value of P becomes smaller, approaching to 0, the querying method is more effective. Based on these experimental results, we can observe that the pruning power of MP_C+ Skyline is better than that of IDC-Index.

Notice that pruning power of a time series dimensionality reduction method is *independent* of the used index structure.

Implemented System. We also need to compare MP_C + Skyline to IDC-Index in terms of implemented systems. The implemented system experiment is evaluated on the normalized CPU cost which is the fraction of the average CPU time to perform a query using the index to the average CPU time required to perform a sequential search. The normalized CPU cost of a sequential search is 1.0.

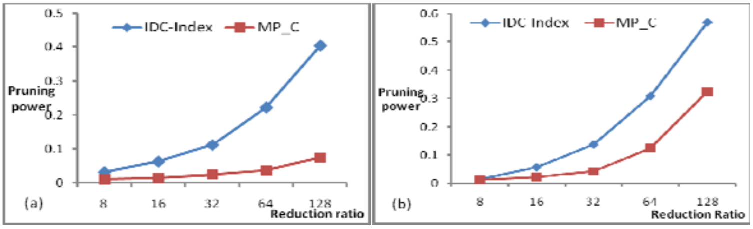


Fig. 3. The pruning powers on Consumer data, tested over a range of reduction ratios (8-128) and query lengths (1024 (a), 512 (b))

The experiments have been performed over a range of query lengths (256-1024), values of reduction ratios (8-128) and a range of data sizes (10000-30000 sequences). For brevity, we show just two typical results. Figure 4 shows the experiment results with a fixed query length 1024.

Between the two competing techniques, in similarity search the MP_C + Skyline index performs faster than IDC-Index based on R*-tree.

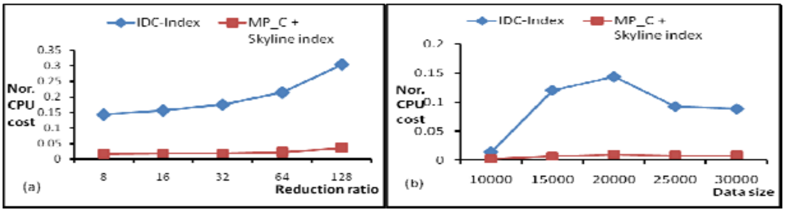


Fig. 4. CPU cost of MP_C using Skyline index and IDC-Index on Consumer data, tested over (a) a range of reduction ratios. (b) a range of data sizes.

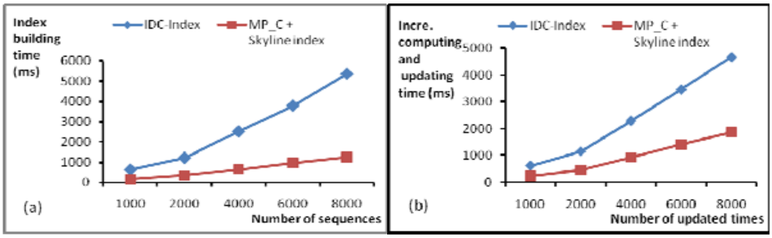


Fig. 5. (a) Index building time and (b) Incremental computing and updating time of MP_C + Skyline and IDC-Index on Consumer data

Building and Updating the Index. We also compare MP_C + Skyline to IDC-Index in terms of the time taken to build the index and the time taken to perform the incremental computation and the delayed update strategy. The experimental results in Figure 5 show that the index building time and the incremental computation plus update time of MP_C + Skyline is lower than that of IDC-Index.

7 Conclusions

We showed that our proposed technique MP_C with the support of Skyline index can be used for the similarity search in streaming time series data. Our approach adopts the same idea of IDC-Index which is based on a incremental computation and a delayed update policy. Experimental results demonstrate that our MP_C method with the support of Skyline index is better than IDC-Index in terms of pruning power and normalized CPU cost. Besides, the index building time along with the incremental computation and update time of MP_C using Skyline index can be faster than IDC-Index based on R*-tree. The limitation in our experiment is that we have not yet adapted the index update rate according to application requirements.

In future, we plan to investigate how to keep the threshold *T* up-to-date as streams evolve with time according to the desirable update frequency. We expect that such threshold could allow number of updates performed to the index which guarantees efficiency.

References

1. Beckman, N., Kriegel, H.P., Schneider, R., Seeger, B.: The R*-tree: An Efficient and Robust Access Method for Points and Rectangles. In: Proc. of 1990 ACM-SIGMOD Conf., Atlantic City, NJ, pp. 322–331 (May 1990)
2. Babu, S., Widom, J.: Continuous queries over data streams. ACM SIGMOD Record 30(3), 109–120 (2001)
3. Guttman, A.: R-trees: a Dynamic Index Structure for Spatial Searching. In: Proc. of the ACM SIGMOD Int. Conf. on Management of Data, June 18-21, pp. 47–57 (1984)
4. Gao, L., Wang, X.: Continually Evaluating Similarity-Based Pattern Queries on a Streaming Time Series. In: Proc. ACM SIGMOD (2002)
5. Kontaki, M., Papadopoulos, A.N., Manolopoulos, Y.: Efficient similarity search in streaming time sequences. In: Proceedings of the 16th International Conference on Scientific and Statistical Database Management (SSDBM 2004), Santorini, Greece (2004)
6. Kontaki, M., Papadopoulos, A.N., Manolopoulos, Y.: Adaptive similarity search in streaming time series with sliding windows. Data & Knowledge Engineering 16(6), 478–502 (2007)
7. Lian, X., Chen, L., Yu, J.X., Wang, G.: Similarity Match over High Speed Time Series Streams. In: Proc. IEEE 23rd International Conference (2007)
8. Liu, X., Ferhatosmanoglu, H.: Efficient k-NN Search on Streaming Data Series. In: Hadzilacos, T., Manolopoulos, Y., Roddick, J., Theodoridis, Y. (eds.) SSTD 2003. LNCS, vol. 2750, pp. 83–101. Springer, Heidelberg (2003)
9. Li, Q., Lopez, I.F.V., Moon, B.: Skyline Index for Time Series Data. IEEE Trans. on Knowledge and Data Engineering 16(6) (2004)
10. Son, N.T., Anh, D.T.: Time Series Similarity Search based on Middle Points and Clipping. In: Proceedings of the 3rd Conference on Data Mining and Optimization (DMO 2011), Putrajaya, Malaysia, June 28-29, pp. 13–19 (2011)

DRFLogitBoost: A Double Randomized Decision Forest Incorporated with LogitBoosted Decision Stumps

Zaman Md. Faisal*, Sumi S. Monira, and Hideo Hirose

Kyushu Institute of Technlogy
680-4 Kawazu, Iizuka, Fukuoka 820-8502, Japan
{zaman,sumi}@ume98.ces.kyutech.ac.jp, hirose@ces.kyutech.ac.jp

Abstract. In this paper, a hybrid decision forest is constructed by double randomization of the original training set. In this decision forest, each individual base decision tree classifiers are incorporated with an additional classifier model, the *Logitboosted* decision stump. In the first randomization, the resamples to train the decision trees are extracted; in the second randomization, second set of resamples are generated from the out-of-bag samples of the first set of resamples. The boosted decision stumps are constructed on the second resamples. These extra resamples along with the resamples on which the base tree classifiers are trained, approximates the original training set. In this way we are utilizing the full training set to construct a hybrid decision forest with larger feature space. We have applied this hybrid decision forest in two real world applications; a) classifying credit scores, and b) short term extreme rainfall forecast. The performance of the hybrid decision forest in these two problems are compared with some well known machine learning methods. Overall results suggest that the new hybrid decision forest is capable of yielding commendable predictive performance.

Keywords: double randomization, decision forest, real logit boosting, credit classification, extreme rainfall prediction.

1 Introduction

The application of collective decision making is prevalent in many research domains, as it often results in a more robust conclusion than any single decision maker could have made alone. In recent decades, the efforts to automate the combination of expert decisions have been studied extensively [18]. The combination of multiple classifiers is known as *classification ensembles*. The fundamental principle is to construct several individual classifiers, and combine them in order to reach a decision (classification) that is more accurate than the individual classifiers [20]. In this way, ensembles improve the generalization performance. Due to this, researchers from several fields including statistical data mining [11], and

[*] Corresponding author.

J.-S. Pan, S.-M. Chen, N.T. Nguyen (Eds.): ACIIDS 2012, Part I, LNAI 7196, pp. 291–300, 2012.

pattern recognition [15] seriously explore the theory and the use of ensemble methodology.

In this paper, a hybrid ensemble method has been proposed, which is composed with two popular classifier ensemble schemes, bagging [1] and adaboost [9] (the most popular) of the Boosting family [21]. The proposed hybrid decision forest ensemble method is motivated by the, "main ideas" of double bagging ensemble [13]. In the novel hybrid decision forest, predictions from simple decision trees and a *boosting* type ensemble are aggregated. In bootstrapping, approximately $\frac{1}{3}$ of the observations of the original training set are left out (defined as, out-of-bag sample (OOBS) by Breiman [2]). In this ensemble, these OOBS are utilized as a separate training set for the adaboost ensemble. The real adaboost [10] is efficient in discarding the course information generated in the adaboost.M1 algorithm [9]. In real adaboost, the predicted labels are transformed into class posteriori probabilities of the outcomes into real valued scale. We have incorporated the binomial log likelihood loss function or the logit loss [10], instead of the exponential loss function of the original adaboost algorithm. The class probabilities of real logitboost with decision stump are utilized to enlarge the feature space of the component base decision tree classifiers of the decision forest. In this way, the decision forest is composed of the usual component decision trees but trained on real logitboost module.

The underlying motivations behind the proposed ensemble are described briefly in the following: a) the double randomization will increase the sparsity of the common instances in each resample to train the decision tree and logitboosted decision stumps. In this way the diversity of the ensemble will increase and after aggregation the variance of the decision tree will be reduced also. b) the adaBoost is highly preferred for 2-class problems, so we have utilized it as the additional classifier model. The exponential loss function of adaboost is non-robust, so the logit loss function (logitboosting) is used, which is robust against noisy problems. In addition to these, it will reduce the bias associated with the construction of the forest; as this type of decision forest incur bias when constructed. The rest of the paper organized as: in Section 2, we have discussed briefly about the constitutional steps of the new decision forest, emphasizing on the real logitboost. In Section 3, we have described the two real world problems handled in this paper with the description of the experiment and discussion of the results of both the problems. This is followed by the Conclusion of the paper.

2 DRFLogitBoost: A Double Randomized Decision Forest Incorporated with LogitBoosted Decision Stumps

In this section, we briefly discuss about the construction of the new decision forest. When a decision tree is adopted as the base learning algorithm, only splits that are parallel to the feature axes are taken into account even though the decision tree is non-parametric and can be quickly trained. Considering that other general splits such as linear ones may produce more accurate trees, a "Double

```
Input:
  – $\mathscr{L}$: training set
  – $\mathscr{X}$: the predictors in the training dataset
  – $B$: number of classifier in the ensemble
  – $\{\omega_1, \dots, \omega_c\}$: the set of class labels
  – $\rho$: small resampling ratio
  – $x$: a data point to be classified

Output: $\omega$: class label for $x$.

Procedure Double Randomized Decision Forest Incorporated with LogitBoosted Decision Stumps ()

  1. For $b = 1, \dots, B$
     (a) $\mathscr{L}^{(b)} \leftarrow$ Resample of size $\rho$ from $\mathscr{L}$.
     (b) $\mathscr{X}^{(b)}$ denote the matrix of predictors $x_1^{(b)}, \dots, x_N^{(b)}$ from $L^{(b)}$.
     (c) $\mathscr{L}^{(2b)} \leftarrow$ Resample from the out-of-bag sample $\mathscr{L}^{(-b)}$ (of size 1-$\rho$).
     (d) $RealLBoost^{(b)} \leftarrow$ A real LogitBoost model constructed on $\mathscr{L}^{(2b)}$.
     (e) $CP^{(b)} \leftarrow$ A matrix with the columns are the class probability of the classes,
         after training $RealLBoost^{(b)}$ on $L^{(b)}$
     (f) $C_{comb}^{(b)} \leftarrow (L^{(b)} \cup CP^{(b)})$ : Construct the combined classifier
     (g) $TCP^{(b)} \leftarrow$ $x$'s class posteriori probability generated by $RealLBoost^{(b)}$.
     (h) $c_{bj}(x, TCP^{(b)}) \leftarrow$ The probability assigned by the classifier $C_{comb}^{(b)}$ that $x$ comes
         from the class $\omega_j$.
     EndFor
  2. Calculate the confidence for each class $\omega_j$, by the "average" combination rule:

         $\mu_j(x) = \frac{1}{B} \sum_{b=1}^{B} c_{bj}((x, TCP^{(b)})), j = 1, \dots, c.$

  3. $\omega \leftarrow$ Class label with the largest confidence.
  4. Return $\omega$
```

Fig. 1. Generic Framework of DRFLogitBoost

Bagging" method was proposed by Hothorn and Lausen [13] to construct ensemble classifiers. In double bagging framework the out-of-bag sample is used to train an additional classifier model to integrate the outputs with the base learning model.

Bagging type ensemble methods reduce the prediction variance by a smoothing operation but without much effect on the bias of the base model. In other words, it is equivalent to say that the bagging type ensemble method with smaller resamples will have poor effect in reducing the bias; but will reduce the variance more than the usual bootstrap sample (size same as the original training set). Though the constitutional steps of this new decision forest is similar to bagging but this is trained on an enlarged feature space. This entails an enhanced representational power for each of the base decision tree, which will lower the bias of the decision forest when combined. Moreover the real logitboost module will produce low biased estimates of class probabilities, which will in a sense increase the efficiency of the additional features for each base tree classifier and hence will increase the prediction accuracy of the decision forest all together.

In our framework each OOBS is randomized once more to construct an real logitboost and then the that real logitboost is applied back to the bootstrap sample to extract the CPP, then all the CPPs of the real logitboost are stored in the matrix CPP and are used as the additional features with the original feature $\mathcal{X}$ as $[\mathcal{X}\ CPP]$ which is an $r \times p(c+1)$, where p is number of features, c is number of classes in D_l. The detailed steps of the proposed method for constructing an decision forest ensemble is described in Fig. 1

– **Input:**

a. X: Training set of size N.
b. C: A classifier, here decision stump.
c. T: Number of classifiers to construct.
d. w: Vector of weights for the observations in X.
e. x: Test instance for classification.
Transform the class labels to 0 and 1. Define this as y^*

1. **Initialization**
Initialize the weights for each observations in the training set as $\frac{1}{N}$.
Also set initial probability $p_0 = 0.5$
let a committee function $F(x) = 0$.
2. **Additive Regression Modeling**
Step 1a. Compute the working response and weights using the probability estimate and class labels as

$$z_t(i) = \frac{y^* - \eta(p_{t-1}(X_i))}{\eta(p_{t-1}(X_i))(1 - \eta(p_{t-1}(X_i)))}$$

$$w_t(i) = \eta(p_{t-1}(X_i))(1 - \eta(p_{t-1}(X_i))$$

where

$$\eta(p) = log\left(\frac{p}{1-p}\right); p \in [0, 1]$$

Step 1b. Fit a regression stump by weighted least square, using these response and weights.
Step 1c. Update $F(X_i)$ by the regression stump.
update the class probability estimate by $\frac{1}{1+\exp{(-2F(X_i))}}$
which is the binomial log-likelihood loss function.
Step 1d. Update the classifier output.
Depending on how many classifier we want to construct, iterate Step 1.
3. **Combining the classifiers**
Step 2. Combine the classifiers using the *weighted majority voting* rule.

Algorithm 1. Real LogitBoosting for 2-class

2.1 Real LogitBoosting (Real LogitBoost)

The adaboost algorithm [9] has been shown empirically to improve classification accuracy. But it produces at each stage, as output the object's predicted label. This coarse information may hinder the efficiency of the algorithm in finding an optimal classification model. To overcome this deficiency, in Friedman et. al [10] proposed an algorithm called Real adaboost. This algorithm outputs a real-valued prediction rather than class labels at each stage of boosting. and the class probability estimate is converted using the half-log ratio to a real valued scale. This value is then used to represent an observation's contribution to the final overall model. Another boosting algorithm developed by Friedman et. al [10], called the LogitBoosting (logitboost from now on) fits an additive logistic regression model by stage-wise optimization of the binomial log-likelihood is more robust in noisy problems where the misclassification risk is substantial. In this paper we have incorporated the real adaboost with binomial log-likelihood loss function, and hence the name, *Real LogitBoosting.* In addition to this, we have used the decision stumps as the base classifier in the logitboost framework. The pseudo-code of real logitboost algorithm is given in Algorithm 1.

3 Problem Setup and Discussion of the Experimental Results

In this section, we have described the two real world problems and later presented the results with the discussion. In both the problems we have employed DRFLogitBoost with 0.30 small resampling ratio and 50 iterations. We have kept the same iterations for all other ensemble methods in this paper to maintain the same magnitude between the algorithms.

3.1 Problem 1: Credit Score Classification

An important issue in financial decision-making is to predict, timely and correctly, business failure e.g. bankruptcy prediction and credit scoring. The credit scoring models permit to discriminate between good credit group and bad credit group. The benefits obtained developing a reliable credit scoring system are [24]:

1. reducing the cost of credit analysis;
2. enabling faster decision;
3. insuring credit collections and diminishing possible risk.

The dataset description is given in Table 1. All these datasets are available at [8].

Experiment Setup Description. We have compared the performance of the hybrid decision forest with several best performing and latest machine learning ensemble classifier methods relevant to credit score classification and binary classification task. We have used, a) random subspace method with neural

Table 1. Description of the 3 Data used in this paper

Dataset	Objects	Classes	Features
Australian Credit	690	2	16
German-credit	1000	2	20
Australian Credit	690	2	15

network (RSM-NN) [17], b) rotation forest (RotF) [19], c) generalized additive model ensemble (GAMens) [4], d) random subspace generalized additive model (GAMrsm) [4] and e) bagged generalized additive model (GAMbag) [4]. We have performed 20 repetitions of 3-CV to compute the following metrics:

1. **Area under the Receiver Operating Characteristic Curve (AUC)**[6]: The Receiver Operating Characteristic curve is a two-dimensional measure of classification performance that plots the True Positive rate against the False Positive rate. In the machine learning literature it is well known that AUC is one of the best method for comparing classifiers in two-class problems [7].
2. **Type I Error** [24]: number of patterns that belong to the bad credit group incorrectly classified into the good credit group.
3. **Type II Error** [24]: number of patterns that belong to the good credit group incorrectly classified into the bad credit group.

In each table for each row (dataset) the best performing method is marked bold. For each table we have performed the comparison of the proposed ensemble method with other classifier methods by two-tailed Wilcoxon Sign Rank test[1]. The significance level is 95%. In all the tables, the method which is significantly worse than the proposed method is marked by a "•" and the method which is significantly better than the proposed method is marked by a "◦".

Discussion of the Results. In Table 2, we have given the AUC values of the ensemble methods stated above. The highest AUC values are marked in bold for each dataset. We see from the table that the new decision forest outperformed most of the methods in two of the datasets (in Australian and Japanese credit data), where as, in German credit data, it is outperformed by the RSM-NN ensemble method.

Table 2. AUC values of the competing algorithms

Dataset	DRFLogitBoost	RSM-NN	RotF	GAMens	GAMrsm	GAMbag
Australian Credit	**0.9408**	0.9338	0.9207 •	0.9126 •	0.9154 •	0.9107 •
German-credit	0.7761	**0.7847**	0.7591	0.7713	0.7729	0.7563 •
Australian Credit	**0.9314**	0.9285	0.9215	0.9178 •	0.9224	0.9006 •

• Significantly worse than DRFLogitBoost at significance level = 95%
◦ Significantly better than DRFLogitBoost at significance level = 95%

[1] For comparing type-I and type-II error in the credit classification task, one sided alternative is employed.

In Table 3, the type-I error values of the ensemble methods are presented. The lowest values are marked in bold for each dataset. It is apparent that, the new ensemble method is not the best performer in the case of lower type-I error. The new decision forest produced lower type-I values than some of the methods in all the datasets. It should be noted that, the type-I values of other methods except the new ensemble method are not consistent. But the new ensemble method has lower type-I error in two datasets and lowest type-I error in one dataset. Considering this, it can be advised that the new ensemble method is better than other methods.

Table 3. Type-I error values of the competing algorithms

Dataset	DRFLogitBoost	RSM-NN	RotF	GAMens	GAMrsm	GAMbag
Australian Credit	10.47	**8.20** ○	15.80 •	13.10 •	9.60	16.86 •
German-credit	**52.72**	66.20 •	53.60	52.91	60.20 •	73.08 •
Australian Credit	11.76	11.60	12.00	14.10 •	**11.60**	15.12 •

• Significantly worse than DRFLogitBoost at significance level = 95%
○ Significantly better than DRFLogitBoost at significance level = 95%

Table 4, represents the type-II error values of the ensemble methods stated above. The best (lowest) values are marked in bold for each dataset. We see from the table, the same pattern of performance of the methods; that is the new decision forest has lower type-II values in two datasets and lowest in one dataset and not any single method produce lower type-II error values for all the datasets. Considering this, the new decision forest can be regarded as a better method than other methods.

Table 4. Type-II error values of the competing algorithms

Dataset	DRFLogitBoost	RSM-NN	RotF	GAMens	GAMrsm	GAMbag
Australian Credit	11.26	17.00 •	**11.20**	12.00	13.62	14.56 •
German-credit	11.49	15.20 •	19.78 •	12.91	**11.32**	23.11 •
Australian Credit	**11.80**	11.89	13.64	12.99	14.80 •	16.42 •

• Significantly worse than DRFLogitBoost at significance level = 95%
○ Significantly better than DRFLogitBoost at significance level = 95%

3.2 Problem 2: Short Term Extreme Rainfall Forecast Problem

In this section, we describe the experimental set-up of short term extreme *categorical* rainfall forecast problem. We have taken the rainfall updates of Fukuoka city of Japan, and compared the forecasting accuracy of the new ensemble with other renowned relevant data-mining techniques : a) LogitBoosting (LB)[5], b) Random Forest (RF) [3], c) Least Square Support Vector Machine (LSSVM)[14].

Dataset Description. In this paper, we have used a dataset containing rainfall records averaged over several weather stations around Fukuoka city of Japan. The dataset contains rainfall update from 1985 to 2009. The available data provides us measures the amount of rainfall (in mm) in a day. In this problem, our aim is to forecast whether or not it will not extreme the next day. We have reconstructed the data using embedding technique. The input variables are the lagged rainfall values of the next day rainfall.

We but we do the next transformation to have a categorical output based on the accumulated rainfall in a day; the categorization is done as follows:

Rainfall amount	Category
0 mm–1 mm	1
2 mm–50 mm	2
50 mm >	3

The first category can also be defined as ‘no rain’ category and the last category can be defined as ‘extreme rainfall’ category. In this paper we are interested to perform the binary forecast of ‘extreme rainfall’ against the ‘normal rain’ event. This is done by defining category 1 and 2 together as ‘normal rain’ event and category 3 is defined as ‘extreme rainfall’ event. We define this forecast problem as ‘Extreme rainfall forecast problem. As the probability of such high amount of rainfall is rear, this forecast problem is about forecasting a rear event. In addition to these the dataset is scaled and normalized before use. We also log transformed all the predictors. We have used linear scaling computed using the training set, to scale the time series. The scaling step is essential to get the time series in a suitable range, between -1, and 1.

Description of Experimental Setup. We have conducted this experiment in two phases, a) Training phase, b) Out of sample (Test) phase. For this purpose, we have partitioned our data in two equal parts, training set and test set. The training period is from 1985 to 1997 and the test phase is from 1998 to 2009.

This is a binary forecast problem, so for forecast verification the usual evaluation metrics are not enough. We have utilized some useful statistics available from signal detection theory which are now frequently used in the climatology [16]. In this article, Mason has provided a universal framework for evaluating the joint probability distribution of forecasts and observations. The computed metrics are fraction correct (FC), hit rate (H), false rate (F) and false alarm ratio (FAR), odds ratio (OR) [23].

In addition to these metrics, specifically for rare event forecasting, the usual metrics (FC, H, F, FAR) can be misleading. Mainly because these measures are heavily influenced by the frequency of the climatological events, so inference on extreme events based on these measures will be rather biased. Stephenson in [23], for verification of rare event forecasting (also extreme event), proposed to use Bias, OR (Odds Ratio), PSS (Pierce Skill Score) and ORSS (Odds ratio skill score). Recently Stepehnson et.al [22] proposed a measure EDS (Extreme Dependency Score) and Hogan in [12] proposed SEDS (Extreme Dependency Score), these two measures are till now quite robust for verification of the rare event forecast. In addition to these we have also computed the AUC (Area under the ROC curve) of models for both the problems. As we know higher AUC is desirable for binary prediction problem and this measure is now used more frequently than accuracy in binary prediction problems.

Discussion of the Results. In Table 5, we have presented the values of the metrics of the prediction algorithms for the extreme rainfall forecasting. The best values of each metric for each of the methods are marked bold (here best is relative to the behaviour of the metrics). It is worth notable that the new decision forest has best skill scores in terms of the forecast verification measures (PSS, EDS) and better values than most other methods for ORSS and SEDS. Also the AUC of the new ensemble is far better than other methods.

Table 5. Metric values of the methods in test phase

Metrics	DRFLogitBoost	LB	RF	LSSVM
FC	0.8980	**0.9162**	0.8897	0.8932
H	0.5000	0.5271	**0.5312**	0.3543
F	0.0083	**0.0056**	0.0282	0.0468
FAR	0.5361	0.4167	**0.2277**	0.5287
Bias	1.540	0.7429	0.7768	0.7411
OR	11.011	**16.182**	11.763	10.704
ORSS	0.8334	**0.8836**	0.8433	0.8291
PSS	**0.4167**	0.2777	0.3193	0.2935
EDS	**0.6332**	0.3891	0.3882	0.3568
SEDS	0.4198	0.388	**0.4651**	0.4459
AUC	**0.8160**	0.5389	0.6596	0.6468

4 Conclusion

In this paper, a new hybrid decision forest is proposed. The decision forest is incorporated with small bootstrap sample aggregation and real logitboosted decision stumps. The decision forest is benefited from the enlarged feature space produced by logitboosted decision stumps for each decision tree during the small bootstrap aggregation. For the performance examination of the new decision forest, it is applied in two real world problems; credit score classification and extreme rainfall forecast. For the performance check, it is compared with some very well known machine learning algorithms available. In the credit classification problem, it is most accurate in classifying the credits and consistent in differentiating between good credits as *good* and bad credits as *bad*. The results of the extreme rainfall prediction suggest that the new decision forest has capability to efficiently perform the categorical rainfall forecasting task.

References

1. Breiman, L.: Bagging Predictors. Machine Learning 24(2), 123–140 (1996)
2. Breiman, L.: Out-of-bag estimation. Tech. Rep. 2 (1996)

3. Breiman, L.: Random Forests. Machine Learning 45(1), 5–32 (2001)
4. De Bock, K.W., Coussement, K., Van den Poel, D.: Ensemble classification based on generalized additive models. Computational Statistics & Data Analysis 54(6), 1535–1546 (2010)
5. Dettling, M., Buhlmann, P.: Boosting for tumor classification with gene expression data (June 2003)
6. Fawcett, T.: An introduction to ROC analysis. Pattern Recognition Letters 27(8), 861–874 (2006)
7. Fawcett, T.: ROC Graphs: Notes and Practical Considerations for Researchers. ReCALL 31(HPL-2003-4), 1–38 (2004)
8. Frank, A., Asuncion, A.: UCI Machine Learning Repository (2010), http://archive.ics.uci.edu/ml
9. Freund, Y., Schapire, R.E.: A Decision-Theoretic Generalization of On-Line Learning and an Application to Boosting. In: Vitányi, P.M.B. (ed.) EuroCOLT 1995. LNCS, vol. 904, pp. 23–37. Springer, Heidelberg (1995)
10. Friedman, J., Hastie, T., Tibshirani, R.: Additive logistic regression: a statistical view of boosting. Annals of Statistics 28(2), 337–407 (2000)
11. Hastie, T., Tibshirani, R., Friedman, J.: The elements of statistical learning: data mining, inference and prediction, 2nd edn. Springer, Heidelberg (2009)
12. Hogan, R.J., O'Connor, E.J., Illingworth, A.J.: Verification of cloud-fraction forecasts. Quarterly Journal of the Royal Meteorological Society 135(643), 1494–1511 (2009)
13. Hothorn, T., Lausen, B.: Double-bagging: combining classifiers by bootstrap aggregation. Pattern Recognition 36(6), 1303–1309 (2003)
14. Suykens, J.A.K., Vandewalle, J.: Least Squares Support Vector Machine Classifiers. Neural Processing Letters, 293–300 (1999)
15. Kuncheva, L.I.: Combining Pattern Classifiers: Methods and Algorithms, 1st edn. Wiley-Interscience (2004)
16. Mason, I.: A model for assessment of weather forecasts. Australian Metereological Magazine 30, 291–303 (1982)
17. Nanni, L., Lumini, A.: An experimental comparison of ensemble of classifiers for bankruptcy prediction and credit scoring. Expert Systems with Applications 36(2), 3028–3033 (2009)
18. Polikar, R.: Ensemble based systems in decision making. IEEE Circuits And Systems Magazine Circuits And Systems Magazine 6(3), 21–45 (2006)
19. Rodríguez, J.J., Kuncheva, L.I., Alonso, C.J.: Rotation forest: A new classifier ensemble method. IEEE Transactions on Pattern Analysis and Machine Intelligence 28(10), 1619–1630 (2006)
20. Rokach, L.: Pattern classification using ensemble methods. World Scientific Publishing (2010)
21. Schapire, R.E.: The Boosting Approach to Machine Learning An Overview. In: MSRI Workshop on Nonlinear Estimation and Classification, vol. 7(4), pp. 1–23 (2003)
22. Stephenson, D.B., Casati, B., Ferro, C.A.T., Wilson, C.A.: The extreme dependency score: a non-vanishing measure for forecasts of rare events. Meteorological Applications 15(1), 41–50 (2008)
23. Stephenson, D.: Use of the "Odds Ratio" for Diagnosing Forecast Skill. Weather Forecasting 15(2), 221–232 (2000)
24. Tsai, C., Wu, J.: Using neural network ensembles for bankruptcy prediction and credit scoring. Expert Systems with Applications 34(4), 2639–2649 (2008)

Learning and Inference Order in Structured Output Elements Classification

Tomasz Kajdanowicz and Przemyslaw Kazienko

Wroclaw University of Technology, Wroclaw, Poland
Faculty of Computer Science and Management
{tomasz.kajdanowicz,kazienko}@pwr.wroc.pl

Abstract. In the paper three learning and inference ordering approaches in the method for structured output classification are presented. As it was previously presented by authors, classification of single element in output structure can be performed by generalization of input attributes as well as already partially classified output elements [9]. The paper addresses crucial problem of how to order elements in the structured learning process to get greater final accuracy. The learning is performed by means of ensemble, boosting classification method adapted to structured prediction - AdaBoostSeq algorithm. Authors present several ordering heuristics for score function application in order to obtain better structured output classification accuracy.

Keywords: Structured Output Prediction, Classifier Ensembles, Multiple Classifier Systems, Boosting.

1 Introduction

A growing interest in the structured output prediction has been observed recently. Machine learning algorithms dealing with structured output prediction problems are able to generalize in a set of training input-output pairs but the input or the output (sometimes both of them) are more complex in comparison with traditional data types. Usually structured prediction works with output space containing complex structure like sequences, trees or graphs. Combined applicability and generality of learning in complex spaces result in a number of significant theoretical and practical challenges.

Considering the standard classification or regression problems, they discover the mapping to output space that is either real-valued or a value from a small, unstructured set of labels. On the other hand structured output prediction depicts the output space that might contain a structure rich in information which can be utilized in the learning process. Due to the profile of structured output prediction it requires in some cases an optimization or search mechanism over complete output space which makes the problem non-trivial.

There exists a variety of structured output prediction problems, e.g. protein function classification, semantic classification of images or text categorization.

J.-S. Pan, S.-M. Chen, N.T. Nguyen (Eds.): ACIIDS 2012, Part I, LNAI 7196, pp. 301–309, 2012.

The structured output prediction (classification) methods solving abovementioned problems might be designed on the basis of two distinct approaches. The former utilizes *a priori* known information about the structure, e.g. makes use of sequential structure and the algorithm tries to benefit from relations between predecessors and consequent. The latter does not assume any shape of output structure and attempts to perform prediction on the pure data.

2 Related Work

The most obvious and directly arising method for structured output prediction is a probabilistic model jointly considering the input and output variables. There exist many examples of such probabilistic models for a variety of inputs and outputs, e.g. probabilistic graphical models or stochastic context free grammars. Given an input, the predicted output might be determined as the result of maximization of the posteriori probability, namely the technique of maximum posteriori estimation. In such approaches, the learning is to model the joint input and output data distribution. However, it is well known that this approach of first modelling the distribution and subsequently using maximum a posteriori estimation for prediction is indirect and might be suboptimal. Therefore, the other, direct discriminative approach might be more appropriate. Such a discriminative learning algorithms perform a prediction on the basis of scoring function optimization over the output space. Recently presented and studied discriminative learning algorithms include Max-margin Markov Nets that consider the structured output prediction problem as a quadratic programming problem [12], a bit more flexible approach that is an alternative adjustment of logistic regression to the structured outputs called Conditional Random Fields [10], Structured Perceptron [1] that has minimal requirements on the output space shape and is easy to implement, Support Vector Machine for Interdependent and Structured Outputs (SVM^{STRUCT}) [14], which applies variety of loss functions and an example of adaptation of the popular back-propagation algorithm - BPMLL [15] where a new error function takes multiple targets into account.

Another approach to structured prediction, LaSO, presented in [2] proposes a framework for predicting structured outputs by learning as search optimization. LaSO allows to reduce the requirement appearing in likelihood- or margin-based algorithms that output structure needs to be computed from the set of all possible structures. As an extension to that approach, SEARN algorithm [3] assumes transformation of structured prediction problems into binary prediction problems to which a standard binary classifier can be applied.

Based on similar assumption another example of structured output algorithm is an extension of the original AdaBoost algorithm to structured prediction. It is the AdaBoostSeq algorithm proposed by authors in [7] and utilized in this paper within experimental studies.

Summarizing, both presented approaches, the generative modelling and the discriminative learning algorithms make use of some scoring function to score each element of the output space. In the case of methods that work in space of partial outputs the order of score function calculation of output space elements is important and may determine the overall accuracy. Therefore appropriate output space learning order is required in order to provide best possible generalization.

3 Structured Output Learning Using AdaBoostSeq Algorithm

In the paper the application of AdaBoostSeq algorithm for structured output prediction is concerned. It is described briefly in this section.

According to the proposal from [3], it is assumed that the structured output classification problem is a cost-sensitive classification problem, where each classification output y_i represents an output structure and is coded as a variable-length vector. The ith data instance is represented by a sequence of T values (T-length sequence): y_i=(y_i^1, y_i^2, . . ., y_i^T) and $y_i^\mu \in C$, where C is a finite set of classes. Additionally, each structure's element $y_i^{\mu_j}$ may be correlated with other elements $y_i^{\mu_k}$, $j \neq k$. Such factorization states the profile of the structure. In AdaBoostSeq algorithm, the modification of the cost function was proposed as well as the new mechanism of learning based on partial output was introduced. AdaBoostSeq algorithm is originally based on the most popular boosting algorithm, AdaBoost [4,13].

The algorithm requires additional assumption that the structured output classification is a binary classification with $y_i^\mu \in \{-1, 1\}$, for $i = 1, 2, \ldots, N$ and $\mu = 1, 2, \ldots, T$, where N is the number of data instances. T is the length of the sequence, i.e. the number of elements in the structure. The general goal of the method is to construct T optimally designed linear combinations of K base classifiers of form $F(x)$:

$$\forall \mu = 1, 2, \ldots, T \quad F^\mu(x) = \sum_{k=1}^{K} \alpha_k^\mu \Phi(x, \Theta_k^\mu) \tag{1}$$

where: $F^\mu(x)$ is the combined, final meta classifier for the μth sequence item (structure element); $\Phi(x, \Theta_k^\mu)$ represents the kth base classifier, performing according to its Θ_k^μ parameters and returning a binary class label for each instance x; α_k^μ is the weight associated to the kth classifier.

Values of the unknown parameters (α_k^μ and Θ_k^μ) are obtained from minimization of prediction error for each μth sequence element for all K classifiers using stage-wise suboptimal method[13]. The actual sequence-loss balancing cost function J is defined as:

$$J(\alpha^\mu, \Theta^\mu) = \sum_{i=1}^{N} \exp(-y_i^\mu(\xi F_{m-1}^\mu(x_i) + (1-\xi) y_i^\mu \hat{R}_m^\mu(x_i) + \alpha^\mu \Phi(x_i, \Theta^\mu))) \tag{2}$$

where: $\hat{R}_m^\mu(x_i)$ is an impact function denoting the influence on prediction according to the quality of preceding sequence labels predictions; ξ is the parameter that allows controlling the influence of impact function in weights composition, $\xi \in \langle 0, 1 \rangle$.

The proposed $\hat{R}_m^{\mu}(x_i)$ impact function is composed as following:

$$\hat{R}_m^{\mu}(x_i) = \sum_{j=1}^{m-1} \alpha_j^{\mu} R^{\mu}(x_i) \tag{3}$$

$$R^{\mu}(x_i) = \frac{\sum_{l=1}^{\mu} y_i^l \frac{F^l(x_i)}{\sum_{k=1}^{K} \alpha_k^l}}{\mu} \tag{4}$$

where: $R^{\mu}(x_i)$ is the auxiliary function that denotes the average coincidence between prediction result $F^l(x_i)$ and the actual value y_i^l weighted with the weights α_k^l associated to the kth base classifiers for all sequence items achieved so far (from 1 to μ) with respect to the value of μ.

The impact function $\hat{R}_m^{\mu}(x_i)$, introduced in Equation 3 and 4, measures the correctness of prediction for all preceding labels $l = 1, \ldots, \mu$ in the sequence. This function is utilized in the cost function and it provides the smaller error deviation for the whole sequence. The greater compliance between prediction and the real value, the higher the function value.

The presented above brief explanation lead to the AdaBoostSeq algorithm for sequence prediction and may be found as full version in [8]. Overall, the AdaBoostSeq algorithm performs the learning of structured output in sequential manner, step by step for each of the structured output elements and results obtained in the previous steps are utilized as an additional input for the following elements. Therefore, the order of elements (the order of learning) may have significant impact on the overall classification accuracy.

As the complete overview over the space of all possible ordering solutions is exponentially complex there should be a method providing reasonable ordering in acceptable time. In order to provide the ordering of the learning steps three approaches formalized as a heuristics in the searching process are proposed.

4 Ordering Heuristics for AdaBoostSeq Algorithm

Searching for the appropriate order of learning steps might be presented as a tree searching problem. In such representation nodes denote particular elements from the output structure. Each level of the tree indicates the appropriate position in the learning order, e.g. the first level of the tree indicates the first position in the learning sequence, whereas each next level corresponds to the following positions in the learning sequence. An example of search tree for structure of four elements is presented in Figure 1 (dashed lines denote example learning order - 2, 3, 4, 1).

The complete overview of the whole tree is highly complex and therefore it is required a less complex method providing approximate, good solution. In order to provide such mechanism, at each of non-leading-to-leaf nodes, the decision of which branch to explore next should be taken. Three heuristics, providing such decisions in the searching process are proposed below.

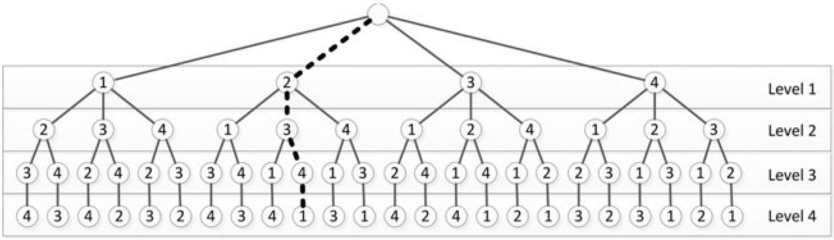


Fig. 1. An example of search tree for structure of four elements composing the learning order for the AdaBoostSeq algorithm

The firsts proposed heuristic (H1) is constructed on the basis of classification error minimization. The heuristic determine which output element should be taken in the next step of learning process. It is done by assessment of the classification error for all remaining elements in the structured output. The element that is classified with the smallest error is taken as next in ordering. Therefore in the structured output of n elements it is needed to perform $\frac{n^2+n}{2} - 1$ evaluations to propose the ordering. For instance if we consider structured output classification of ten elements the heuristics will require ten error evaluations to determine the first element in the learning order. The second element will be chosen from nine and so on, up to the tenth element, together 54 evaluations.

The second proposed heuristic (H2) provides similarly the decision of exploration on the basis of minimum classification error. The output element that should be taken in the next step of learning process is obtained by accessing the classification error obtained in the second level descendant node. For instance, discovering the element to be chosen on the level 1 (first element in the learning sequence), the heuristic accedes all descendant nodes (level 2) and compute the classification error using input attributes enriched with the first chosen output element. The element at level 1 is chosen when it provides learning ability resulting in the smallest average error for all descendants (at level 2). The heuristics computing the output structure of n elements requires to perform $\frac{n^3-n}{3} - 1$ evaluations to propose the ordering.

The third heuristic (H3) is realized likewise the second one (H2), but the decision of the node to be chosen as the next learning element is taken by computing average classification accuracy for the pair of elements: parent(ascendant) and child (descendant). The node which has minimal average classification error with any of its descendant nodes is chosen then. This heuristics requires $\frac{n^3}{3} + \frac{n^2}{2} + \frac{n}{6} - 1$ evaluations for the structure of n elements to provide an ordering.

5 Experiments and Results

The main objective of the performed experiments was to evaluate the classification accuracy of the AdaBoostSeq algorithm driven by various learning ordering of output elements. The method was examined according to Hamming Loss and

Classification Accuracy for six distinct datasets. Some standard evaluation measures from the previous work have been used in the experiments. The utilized measures are calculated based on the differences of the actual and the predicted sets of labels (classes) over all cases x_i in the test set. The first measure is Hamming Loss HL, which was proposed in [11] and is defined as:

$$HL = \frac{1}{N} \sum_{i=1}^{N} \frac{Y_i \triangle F(x_i)}{|Y_i|} \tag{5}$$

where: N is the total number of cases x_i in the test set; Y_i denotes actual (real) labels (classes) in the sequence, i.e. entire structure corresponding to instance x_i; $F(x_i)$ is a sequence of labels predicted by classifier and $\triangle$ stands for the symmetric difference of two vectors, which is the vector-theoretic equivalent of the exclusive disjunction in Boolean logic.

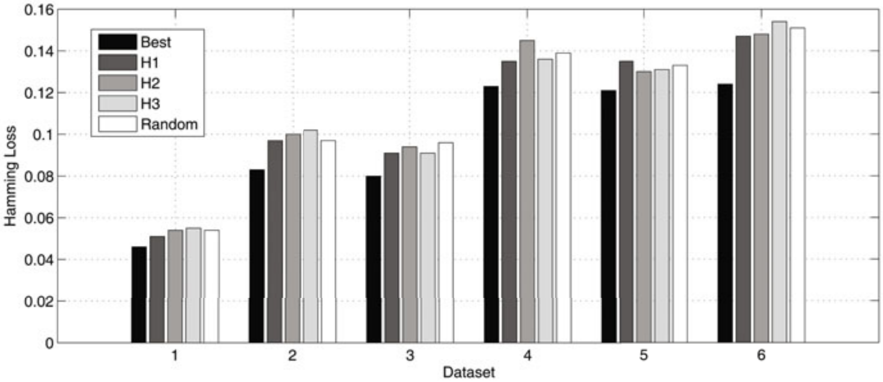


Fig. 2. Hamming Loss error of examined ordering methods (Best - ordering with the smallest Hamming Loss error, H1, H2, H3 - proposed heuristics, Random - random ordering)

The second evaluation measure utilized in the experiments is Classification Accuracy CA [5], defined as follows:

$$CA = \frac{1}{N} \sum_{i=1}^{N} I(Y_i = F(x_i)) \tag{6}$$

where: $I(true) = 1$ and $I(false) = 0$,

Measure CA is a very strict evaluation measure as it requires the predicted sequence of labels to be an exact match of the true set of labels.

The main attention in the experiment was concentrated on the evaluation of three proposed heuristics deriving distinct ordering schemes compared to the best possible learning order (obtained in complete overview of ordering permutations) and the random one (obtained randomly from uniform distribution).

The performance of the analysed methods was evaluated using 10-fold cross-validation and the evaluation measures from Equation 5 and Equation 6.

The experiments were carried out on six distinct, real datasets from the same application domain of debt portfolio pattern recognition [6]. Datasets represent the problem of aggregated prediction of sequential repayment values over time for a set of claims. The structured output for each debt (case) is composed of a vector of binary indicator denoting whether at a certain period of time it was repaid at a certain level. The output to be predicted is provided for all consecutive periods of time the case was repaid. For the purpose of the experiment, the output was limited to only six elements. The number of cases in the datasets varied from 1,703 to 6,818, while the number of initial, numeric input attributes was the same: 25. Note that the number of input attributes refers only classification for the first element in the sequence; for the others, outputs of the preceding elements are added to their input.

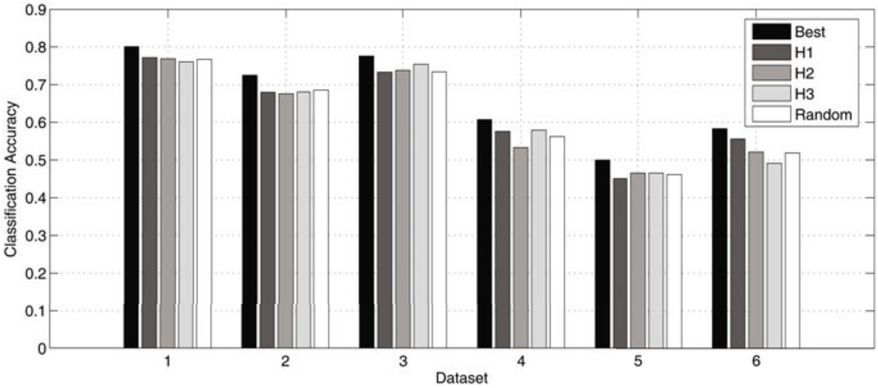


Fig. 3. Classification Accuracy of examined ordering methods (Best - ordering with the best classification accuracy, H1, H2, H3 - proposed heuristics, Random - random ordering)

The results of Hamming Loss and Classification Accuracy are presented in Table 1, Figure 2 and 3. Its relative values, with respect to the best possible order (Best), i.e. $(x - Best)/Best \cdot 100\%$, are depicted in Table 2.

As it was discovered by preliminary experimentation the average differences between the worst possible and the best possible results depending on the learning order are about 30% (for HL) and 14% (for CA). This is the margin, within which proposed heuristic methods can be searched. The heuristic method H1 provided the ordering resulting in usually better than other ordering methods (H2, H3, Random). In particular, this heuristic approach usually gains in both CA and HL compared to the other orderings and appears to be a rational solution as it has the lowest complexity.

Table 1. Results obtained in the experiments, where orderings denote: Best - ordering with the best classification accuracy, H1, H2, H3 - proposed heuristics, Random - random ordering; HL - Hamming Loss, CA - Classification Accuracy

	Ordering method									
	Best		H1		H2		H3		Random	
Dataset	HL	CA	HL	CA	HL	CA	HL	CA	HL	CA
1	0.046	0.801	0.051	0.772	0.054	0.7692	0.0547	0.7608	0.054	0.767
2	0.083	0.725	0.097	0.680	0.0995	0.6762	0.1024	0.6806	0.097	0.686
3	0.080	0.776	0.091	0.733	0.0942	0.7375	0.0907	0.7535	0.096	0.734
4	0.123	0.607	0.135	0.576	0.1447	0.5327	0.1356	0.5791	0.139	0.562
5	0.121	0.500	0.135	0.451	0.1301	0.4653	0.131	0.4648	0.133	0.461
6	0.124	0.583	0.147	0.556	0.1481	0.5205	0.1536	0.4912	0.151	0.519
Average	0.096	0.665	0.109	0.628	0.112	0.617	0.111	0.622	0.112	0.621

Table 2. Average relative errors in comparison with the best ordering, in percentage (0 denotes the same accuracy like obtained with the best ordering), the orderings are: Best - ordering with the best classification accuracy, H1, H2, H3 - proposed heuristics, Random - random ordering; HL - Hamming Loss, CA - Classification Accuracy

	Ordering method							
	H1		H2		H3		Random	
Dataset	HL	CA	HL	CA	HL	CA	HL	CA
1	10,66%	3,54%	15,47%	3,93%	16,77%	4,97%	16,10%	4,25%
2	14,16%	6,27%	16,47%	6,76%	19,39%	6,16%	13,60%	5,47%
3	10,90%	5,47%	14,83%	4,93%	11,11%	2,87%	16,72%	5,34%
4	8,60%	5,15%	15,09%	12,28%	8,80%	4,64%	11,26%	7,38%
5	10,40%	9,78%	6,71%	6,94%	7,40%	7,04%	8,68%	7,86%
6	15,22%	4,69%	16,09%	10,75%	19,81%	15,77%	18,02%	11,04%
Average	11,72%	5,61%	13,88%	7,28%	13,50%	6,56%	13,75%	6,60%

6 Conclusions

The problem considered in the paper concerns discovering appropriate learning order in the structured output prediction. It was based on AdaBoostSeq algorithm that assumes that labels of the already classified output elements are used as additional input features for the next elements. Since the elements in the sequence may be correlated, the order of learning may influence accuracy of the entire classification. According to experiments' results, the margin between the worst and the best order may be even several dozen of percent for Hamming Loss measure and for Classification Accuracy. Moreover, three proposed heuristics, each of distinct complexity, showed the ability to result in better than random learning order. Overall, the H1 heuristic method proposed in the paper appears to be a reasonable direction to find the learning order providing better results than the other simple orders. It is at the same time much less computationally expensive than checking all possible orders to find the best one.

Further research will be focused on extended studies on the properties of proposed heuristics as well as development of some other heuristic ordering methods.

Acknowledgement. This work was partially supported by The Polish Ministry of Science and Higher Education the research project 2011-2012, 2011-2014 and Fellowship co-financed by The European Union within The European Social Fund.

References

1. Collins, M.: Discriminative training methods for hidden Markov models: Theory and experiments with perceptron algorithms. In: Conference on Empirical Methods in Natural Language Processing, vol. 10, pp. 1–8 (2002)
2. Daume, H., Marcu, D.: Learning as Search Optimization: Approximate Large Margin Methods for Structured Prediction. In: International Conference on Machine Learning, ICML 2005 (2005)
3. Daume, H., Langford, J., Marcu, D.: Search-based structured prediction. Machine Learning 75, 297–325 (2009)
4. Freund, Y., Schapire, R.: A decision-theoretic generalization of on-line learning and an application to boosting. Journal of Computer and System Sciences 55, 119–139 (1997)
5. Ghamrawi, N., McCallum, A.: Collective multi-label classification. In: Proceedings of the 3005 ACM Conference on Information and Knowledge Management, pp. 195–200 (2005)
6. Kajdanowicz, T., Kazienko, P.: Prediction of Sequential Values for Debt Recovery. In: Bayro-Corrochano, E., Eklundh, J.-O. (eds.) CIARP 2009. LNCS, vol. 5856, pp. 337–344. Springer, Heidelberg (2009)
7. Kajdanowicz, T., Kazienko, P.: Boosting-based Sequence Prediction. New Generation Computing 29(3), 293–307 (2011)
8. Kazienko, P., Kajdanowicz, T.: Base Classifiers in Boosting-based Classification of Sequential Structures. Neural Network World 20, 839–851 (2010)
9. Kajdanowicz, T., Kazienko, P.: Structured Output Element Ordering in Boosting-Based Classification. In: Corchado, E., Kurzyński, M., Woźniak, M. (eds.) HAIS 2011, Part II. LNCS (LNAI), vol. 6679, pp. 221–228. Springer, Heidelberg (2011)
10. Lafferty, J., McCallum, A., Pereira, F.: Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In: International Conference on Machine Learning, ICML 2001, pp. 282–289 (2001)
11. Schapire, R.E., Singer, Y.: Boostexter: a boosting-based system for text categorization. Machine Learning 39, 135–168 (2000)
12. Taskar, B., Guestrin, C., Koller, D.: Max-margin Markov networks. In: Advances in Neural Information Processing Systems, vol. 16, pp. 25–32. MIT Press, Cambridge (2004)
13. Theodoris, S., Koutroumbas, K.: Pattern Recognition. Elsevier (2009)
14. Tsochantaridis, I., Hofmann, T., Thorsten, J., Altun, Y.: Large margin methods for structured and interdependent output variables. Journal of Machine Learning Research 6, 1453–1484 (2005)
15. Zhang, M.L., Zhou, Z.H.: Multi-label neural networks with applications to functional genomics and text categorization. IEEE Transactions on Knowledge and Data Engineering 18, 1338–1351 (2006)

Evaluating the Effectiveness of Intelligent Tutoring System Offering Personalized Learning Scenario[*]

Adrianna Kozierkiewicz-Hetmańska

Institute of Informatics, Wroclaw University of Technology, Poland
adrianna.kozierkiewicz@pwr.wroc.pl

Abstract. In this paper the prototype of an e-learning system that incorporates learning style is described. This system with a personalized courseware was used to conduct an experiment. The e-learning system collected information about a student during the registration process. Next, a user was assigned to an experimental or a control group. Depending on the previous classification the student was offered a learning scenario suited to individual learner's preferences or a universal learning scenario. The research was devoted to measure students' learning results from both groups. Significantly higher results were obtained by learners whose user profiles were taken into consideration during the determination of the learning scenario.

1 Introduction

Intelligent tutoring systems also called e-learning systems or systems for distance education are very popular among teachers and students. Those systems allow a teacher to save time because once prepared educational materials could be used many times by many students and prepared tests do not need laborious assessment. Moreover, it is possible to work out learning material in an interesting way by applying various multimedia techniques. Students are allowed to learn at any convenient time and place using e-learning systems. Furthermore, intelligent tutoring systems increase the learning effectiveness: users achieve better learning results in a shorter time.

In order to keep the intelligent tutoring systems interesting they require further development and detailed analysis. It is also needed to conduct a research on influence of the applied didactic methods on effectiveness of the learning process. Moreover, the implemented and prepared systems should be evaluated by users. It is one of the most important step of a system design which is very often omitted. So far there are not enough works containing information about quality of designed e-learning systems confirmed by experimental research or it is not possible to generalize conclusions because they concern only systems which have been tested.

In this work we want to introduce the results of experiments which prove the hypothesis that *it is possible to increase the effectiveness of intelligent tutoring systems by taking into consideration the user profile in determination of the learning scenario.*

[*] This research was financially supported by Polish Ministry of Science and Higher Education under the grant N N519 407437.

J.-S. Pan, S.-M. Chen, N.T. Nguyen (Eds.): ACIIDS 2012, Part I, LNAI 7196, pp. 310–319, 2012.

By the learning scenario we mean the order and the presentation form of an educational material proposed to a student by the e-learning system. The formal definition of the learning scenario can be found in [8], [11],[12]. It is intuitively true that students are interested and more motivated if they learn using intelligent tutoring systems where the learning materials and tests are suitable for students' preferences, needs, interests or current knowledge level. Our research confirms the efficiency of e-learning systems with personalization of the learning process. In addition, we check the influence of personalization of the learning process according to age.

We designed and implemented a prototype of an e-learning system which was used to conduct an experiment. The prototype of the e-learning system is a simplification of an intelligent tutoring system proposed in author's previous work. The author designed the model of the intelligent tutoring system which offers an individual learning process on each step. After the registration process the student is classified to a group of similar users based on a set of attributes selected by an expert as a criterion of the classification. The opening learning scenario for a new learner is determined based on finished scenarios of students who belong to the same class as the new one. For this task the algorithms using the consensus theory are worked out [8], [12], [13]. After each lesson the user has to pass a computer adaptive test where each question is selected based on the current student's knowledge level [10]. If the student achieves a sufficient test score (more than 50%) he continues learning according to the previously selected learning scenario. Otherwise, it is a signal for the system that the opening learning scenario should be modified. The author proposed two methods which have been described in [9], [11], [12] and [13]. The learning process is finished if all lessons from the learning scenario are taught.

In the next Section an overview of previous researches concerning assessment of methods applied in various intelligent tutoring systems is presented. Section 3 describes the concept of an experiment with a short description of the implemented e-learning system prototype. Section 4 shows the results of an experiment with appropriate conclusions. Finally, general conclusions and future work are described.

2 Related Works

The evaluation of intelligent tutoring systems is an important but often neglected stage of an e-learning system development. In many works it is possible to find formal system models for distance education without an experimental proof of efficiency of those system. Sometimes researches are limited to a system which has been tested and conclusions are not general or the results are not statistically significant.

System ELM-ART [18] is one of the first and adaptive web-based system where various methods are applied. In ELM-ART knowledge is represented by means of the multi-layered overlay model which supports adaptation in the system. System offers individual curriculum sequencing and adaptive annotation of links. Additionally, the system supports adaptive testing and stores student's preferences which are used to create a learning environment suitable for user's needs. System ELM-ART was evaluated only in comparison to previous version of this system called ELM-PE. Moreover, the research was conducted on a small population (less than 30). Tests showed that system ELM-ART is better than ELM-PE.

EFit [6] is a system which provides highly individualized instructions and teaching environment adapted to learner's knowledge and learning capabilities. Authors presented a study carried out with 194 children at a German lower secondary school. The results showed that, compared to a non-treatment control group, children improved their arithmetic performance if they learned using eFit.

In Shikshak [4] knowledge is represented as a tree or a topic dependency graph. Student model stores the performance of a learner represented by a fuzzified value and some additional information. Shikshak plans the individual teaching method based on user's performance level and selects the type of material (such as explanation based, visual based, instructional) as suited to the learning style of a particular student using fuzzy rules. The study based on Shikshak system were conducted on a small population (only thirty three students). The results demonstrated an overall 4% improvement in performance while teaching using Shikshak in comparison with the performance of teaching in traditional classrooms. Additionally, authors presented some inherent features of the system. The received results concerned only tested system.

SEDHI [17] classifies the students according to profiles (defined based on a statistical study of the course user and usage data) and adapts the navigation using the techniques of link hiding and link annotation. The evaluation of the prototype allowed only to demonstrate the appropriateness of applied methods but results were not binding because only five students took part in the experiment.

DEPTHS [7] is another intelligent tutoring system which generates teaching plans and adaptive presentation of the teaching material. DEPTHS also supports adaptive presentation and adaptive navigation. In this system knowledge is organized in a dependency graph. The student model stores and updates data about user's performance within a specific subject domain. The efficiency of DEPHTS was evaluated by 42 students. The first assessment relied on measuring the learner's satisfaction level. Students reported that the system had helped them to learn and provided many useful information, feedback messages and advice for further work. Authors demonstrated that students who learned with DEPTHS performed better than students who learned in the traditional way. Finally, the experiment found that the proposed student model did not reflect the realistic students' knowledge.

System WELSA [16] tries to identify students' learning preferences and based on worked out adaptation rules offers students individualized courses. Authors of WELSA involve 64 undergraduate students in an experiment. Users could follow two courses (in the Computer Science area): one adaptive and one non-adaptive. The obtained results show that the matched adaptation approach increased the efficiency of the learning process with a lower amount of time needed for studying and a lower number of hits on learning resources.

More general studies are presented in [3]. System AnimalWatch [3] offers problems and tasks customized to student's proficiency level and adaptively focuses on the areas that each student needs to practice most. Authors of this system present researches involved pre and post test comparison. Results indicated that students improved from pre to post test after working with AnimalWatch. Improvement was greatest for students with the weakest initial math skills who were also most likely to use the multimedia lessons and worked examples. Authors studied over 400 students attending schools in Los Angeles who used AnimalWatch.

Pillay and Wilss [15] pointed out that students achieve better results if the material was matched to individuals' preferred cognitive style. Students tended to perform better (66%) than those who received mismatched instruction (62%). The small sample size (only 26 students) and the uneven distribution over the cognitive styles prevented the computation of any statistically significant analyses.

Another study was carried out by Kwok and Jones [14]. Authors found that students at the extremes of the learning style spectrum needed guidance in selecting an appropriate navigation method and it helped raise their interest in the material.

In paper [1] the results of 3 experiments were introduced. The studies demonstrated that both introverts and extroverts had benefits from adaptive e-learning systems but extroverted learners would benefit even more from using those systems. The sample size of the experiments (less than 40) could not guarantee the validity of the interpretations.

The experiment conducted as a part of this study is most similar to experiment described in paper [2] where authors used the Felder-Solomon Learning Style Questionnaire to measure the learning style preferences of students. The obtained results show that all students achieved significantly higher scores while browsing session matched to their learning style. However, in our research we examined bigger population (more than 21 people) and learning material was prepared for all learning styles (not only global and sequential). Additionally, we investigate the influence of personalization of the learning process according to age.

3 The Concept of an Experiment

In our researches we want to show that students achieve better learning results if the learning process is personalized to student's preferences and needs. For this job the prototype of an e-learning system was implemented. The intelligent tutoring system was prepared based on content management system Joomla and its plug-in VirtueMart (with several custom modifications). The e-lessons were created using HTML and CSS and tests were created using Hot Potatoes. The results of experiment were stored in MySQL database.

Users who take part in the experiment provide information about themselves in the first step. The system collects information about student's e-mail, login, password, age, sex, educational level, learning style according to Felder's and Silverman's theory [5] and if they have a driving license. Users could be assigned to two groups: a control group and an experimental group based on sex, educational level, age and possession of a driving license. Both groups should have similar distribution of women, man, young people, old people, middle-age people, educated people, uneducated people, with and without a driving license. If a student is assigned to a control group he is offered a universal learning scenario the same as other students in that group. The universal learning scenario is a combination of dedicated learning scenarios and is composed of some pictures, text, films, tasks etc. If the user is classified to an experimental group he is presented with educational materials suitable for his learning style. The learning scenarios the most appropriate for each student's learning style were designed based on the teaching guidelines found in the literature. After detailed analysis of learning styles, 11 learning scenarios were distinguished which differ from each other in lessons' order and presentation methods. Students described as global were proposed the learning scenario with an additional presentation at the beginning, which contains the big picture and the learning material is organized from general ideas to details. Sequential students were presented the material logically ordered from details to general ideas. The learning scenario for visual students contained more pictures, flows, charts etc. Verbal students got more out of words. Active learners were offered more tests and practical tasks [5]. The author implemented, within the prototype of e-learning system, the rules for tailoring the learning scenario to the student's characteristic.

For all users the learning material is about intersections, roadway signs related to intersections and right-of-way laws. After learning, the user has to pass a test within 10 minutes. The test consists of 10 questions chosen randomly from a question bank consisting of 30 items. After solving the test the learner is presented with a test score in percentage. If the student fails the test (more than 50% of wrong answers) for the first time he is offered the same learning scenario once again. After the second failure the system chooses the learning scenario with students achieving the best results. If the test score is still unsatisfactory the experiment is finished without a successful graduation.

4 The Results of Experiment

The research was conducted using the prototype of an e-learning system described in Section 3. In our experiment there were 297 participants. In this group 123 persons did not have any driving licenses which makes it highly probable that they have learned the presented learning material for the first time. The aim of our research was to verify the hypothesis that students who are offered the personalized learning material (from the experimental group) achieve better results than students who are presented the universal learning scenario (from the control group). The results of the described experiment are presented in Figure 1.

Hypothesis 1. The mean test scores of experimental and control groups for students without driving licenses satisfies the following assumption: $H_0 : \overline{\mu}_{exs} = \overline{\mu}_{cont}$ or alternative assumption: $H_1 : \overline{\mu}_{exs} > \overline{\mu}_{cont}$.

Proof. Firstly, we check the normality of the distribution of analyzed features using the Shapiro-Wilk test. For both groups (experimental and control) we can conclude the normality of distribution of analyzed features ($W_{\exp} = 0.9516 > W_{(0.05,55)} = 0.951$ and $W_{cont} = 0.962 > W_{(0.05,67)} = 0.956$) and next, we use the parametric test. There is no information considering standard deviation and sizes of groups are different but large ($n_{exs} + n_{cont} \geq 120$) that is why for testing the null hypothesis the statistic follow normal distribution $N(0,1)$ is assumed:

$$u = \frac{\overline{\mu}_{exs} - \overline{\mu}_{cont}}{\sqrt{\frac{S_{exs}{}^2}{n_{exs}} + \frac{S_{cont}{}^2}{n_{cont}}}} \qquad (1)$$

where: $\overline{\mu}_{exs}$ - average score value for the experimental group, $\overline{\mu}_{cont}$ - average score value for the control group, S_{exs} - estimated standard deviation of the sample for the experimental group, S_{cont} - estimated standard deviation of the sample for the control group, n_{exs} -size of the experimental group, n_{cont} -size of the control group. The tested statistical value equal:

$$u = \frac{59.818 - 52.239}{\sqrt{\frac{(23.08)^2}{55} + \frac{(22.248)^2}{67}}} = 1{,}834$$

The significance level of 0.05 is assumed. The confidence interval equals $[1.64, +\infty)$. The tested statistical value belongs to the confidence interval $u \in [1.64, +\infty)$, that is why the null hypothesis is rejected and the alternative hypothesis is assumed.

We can estimate the difference between the average test scores of experimental and control group. Let us determine the following confidence interval:

$$((\overline{\mu}_{exs} - \overline{\mu}_{cont}) - \frac{u_{0.95}}{\sqrt{\frac{S_{exs}{}^2}{n_{exs}} + \frac{S_{cont}{}^2}{n_{cont}}}}; (\overline{\mu}_{exs} - \overline{\mu}_{cont}) + \frac{u_{0.95}}{\sqrt{\frac{S_{exs}{}^2}{n_{exs}} + \frac{S_{cont}{}^2}{n_{cont}}}}) \qquad (2)$$

We obtain: $(7.182; 7.976)$

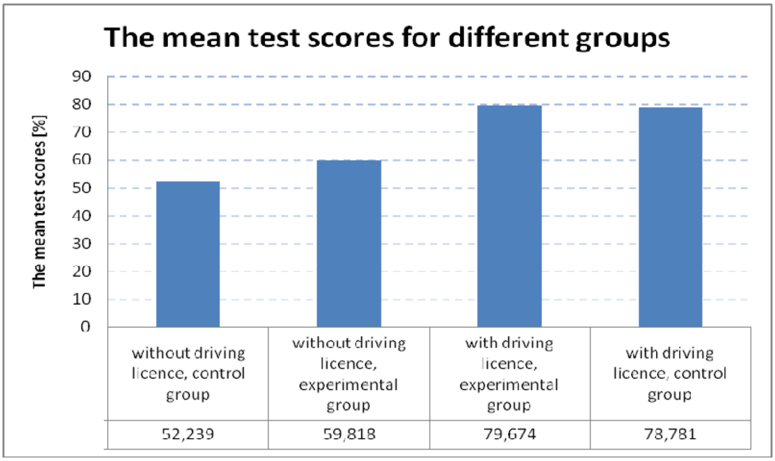


Fig. 1. The mean test scores for different groups

Conclusion 1. *The mean test score of the experimental group is greater than the mean score of the control group in case that students have no driving licenses. Students who were offered the personalized learning scenario achieve better learning results by more than 7.182% and less than 7.976% than students who were proposed the universal learning scenario.*

Hypothesis 2. The mean test scores of experimental and control groups for students with driving licenses satisfies the following assumption: $H_0 : \overline{\mu}_{exs} = \overline{\mu}_{cont}$ or alternative assumption: $H_1 : \overline{\mu}_{exs} > \overline{\mu}_{cont}$.

Proof. For both groups (experimental and control) we cannot reject the hypothesis about the normality of distribution of analyzed features (the Shapiro-Wilk test: $W_{\exp} = 0.9713 > W_{(0.05,92)} = 0.963$ and $W_{cont} = 0.965 > W_{(0.05,82)} = 0.961$). As it was done before we calculate the statistic (2):

$$u = \frac{79.674 - 78.781}{\sqrt{\frac{(16.65)^2}{92} + \frac{(17.49)^2}{82}}} = 0.34$$

The significance level of 0.05 is assumed. We have $u \notin [1.64, +\infty)$, that is why the null hypothesis cannot be rejected.

Conclusion 2. *The mean test score of experimental and control groups are not significantly different in case that students have a driving license.*

Additionally, we investigate how the personalization of learning material influence the efficiency of learning process according to age. The results are presented in Figure 2 only for people who did not a have driving license.

Hypothesis 3. The mean test scores of experimental and control groups for users under age of 18 satisfies the following assumption: $H_0 : \overline{\mu}_{exs} = \overline{\mu}_{cont}$ or alternative assumption: $H_1 : \overline{\mu}_{exs} > \overline{\mu}_{cont}$.

Proof. Firstly, we verify the normality of the distribution of analyzed features by using the Shapiro-Wilk test. For both groups (experimental and control) we can accept the normality of the distribution of analyzed features ($W_{\exp} = 0.9764 > W_{(0.05,16)} = 0.887$ and $W_{cont} = 0.9188 > W_{(0.05,24)} = 0.916$). There is no information considering the standard deviation and sizes of groups are also different. We need to check the equality of the standard deviation using F-Snedecor test. We obtain $F = 1.586$. For $\alpha = 0.05$ the tested statistical value does not belong to the confidence interval $[2.47, +\infty)$ (two-tailed test), so we cannot reject the hypothesis that the standard deviation of two samples are equal. For testing the null hypothesis the t-Student test is assumed:

$$t = \frac{\overline{\mu}_{exs} - \overline{\mu}_{cont}}{\sqrt{\frac{(n_{exs} - 1) \cdot S_{exs}{}^{2} + S_{cont}{}^{2} \cdot (n_{cont} - 1)}{n_{exs} + n_{cont} - 2} \cdot \left(\frac{1}{n_{exs}} + \frac{1}{n_{cont}} \right)}} \tag{3}$$

The statistic is equal to:

$$t = \frac{48.75 - 35.417}{\sqrt{\frac{(16-1) \cdot (20.879)^{2} + (16.578)^{2} \cdot (24-1)}{16+24-2} \cdot \left(\frac{1}{16} + \frac{1}{24} \right)}} = 2.245$$

The significance level of 0.05 is assumed. The tested statistical value belongs to the confidence interval $t \in [1.686, +\infty)$, that is why the null hypothesis is rejected and the alternative hypothesis is assumed.

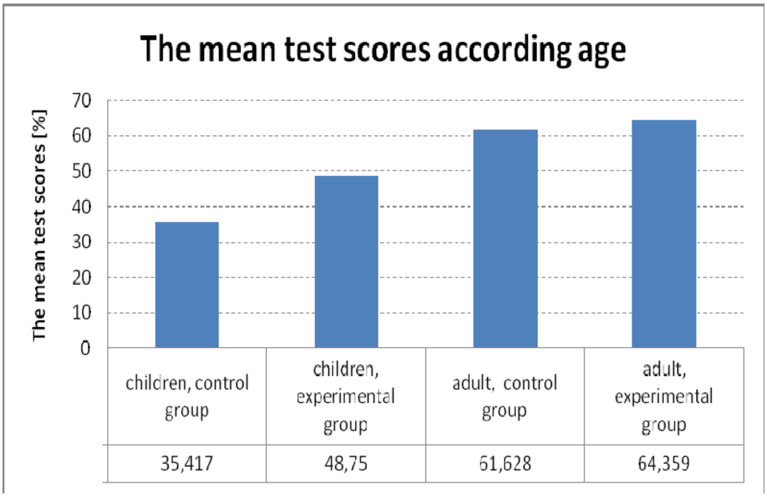


Fig. 2. The mean test scores according age

Conclusion 3. *The mean test score of the experimental group is greater than the mean score of the control group in case that users are underage.*

Hypothesis 4. The mean test scores of experimental and control groups for users over age of 18 satisfies the following assumption: $H_0 : \overline{\mu}_{exs} = \overline{\mu}_{cont}$ or the alternative assumption: $H_1 : \overline{\mu}_{exs} > \overline{\mu}_{cont}$.

Proof. In both groups (experimental and control) we cannot reject the hypothesis about the normality of the distribution of analyzed features (the Shapiro-Wilk test: $W_{\exp} = 0.9438 > W_{(0.05,39)} = 0.939$ and $W_{cont} = 0.949 > W_{(0.05,43)} = 0.943$). As it was done before we check the equality of the standard deviation using the F-Snedecor test. We obtain $F = 1.349$. For $\alpha = 0.05$ the tested statistical value does not belong to the confidence interval $[1.88,+\infty)$ (two-tailed test), so we cannot reject the hypothesis that the standard deviation of two samples are equal. Next, we calculate the statistic (3):

$$t = \frac{64.359 - 61.628}{\sqrt{\frac{(39-1)\cdot(22.395)^2 + (19.281)^2\cdot(43-1)}{39+41-2}\cdot\left(\frac{1}{39}+\frac{1}{41}\right)}} = 0.579$$

For the significance level equal to 0.05 we obtain the following confidence interval: $[1.664,+\infty)$. The tested statistical value does not belong to the confidence interval $t \notin [1.664,+\infty)$, that is why the null hypothesis cannot be rejected.

Conclusion 4. *The mean test score of experimental and control groups are not significantly different in case that students are adults.*

5 Conclusion and Further Work

The personalization of the learning scenario increases the effectiveness of the learning process: the mean test score of the experimental group is greater than the mean test score of the control group. That conclusion is only true for students who are probably presented with the learning material for the first time. Users who have learned the material before do not use the proposed material but solve the tests based on the previously acquired knowledge. Taking into account student's preferences and the learning style improves the learning results by more than 7.182% and less than 7.976%. Furthermore, the personalization have the statistically significant influence on the improvement of the learning result in case that users are underage.

The results of the conducted research should be used by creators of intelligent tutoring systems. The designed e-learning systems should be able to offer an educational material suitable for user's learning style, knowledge, interests, abilities etc. and the individualized learning process on each step.

In the further work we are planning to collect and work out more experimental results. We want to show an influence of the personalization of the learning process according to sex and education level. We want to investigate which learning scenario was the most popular and which was giving the best results.

References

1. Amal Al-Dujaily, A., Ryu, H.: A Relationship between e-Learning Performance and Personality. In: Proceedings of the Sixth IEEE International Conference on Advanced Learning Technologies, Kerkrade, Holandia, pp. 84–86 (2006)

2. Bajraktarevic, N., Hall, W., Fullick, P.: Incorporating learning styles in hypermedia environment: Empirical evaluation. In: Proceedings of AH 2003, at the 12th World Wide Web Conference, Budapest, Hungary, pp. 41–52 (2003)
3. Beal, C.R., Arroyo, I.M., Cohen, P.R., Woolf, B.P.: Evaluation of AnimalWatch: An intelligent tutoring system for arithmetic and fractions. Journal of Interactive Online Learning 9, 64–77 (2010)
4. Chakraborty, S., Roy, D., Basu, A.: Shikshak: An Architecture for an Intelligent Tutoring System. In: Proc. 16th International Conference on Computers in Education, Taipei, Taiwan, pp. 24–31 (2008)
5. Felder, R.M., Silverman, L.K.: Learning and Teaching Styles in Engineering Education. Engineering Education 78(7), 674–681 (1988)
6. Graff, M., Mayer, P., Lebens, M.: Evaluating a web based intelligent tutoring system for mathematics at German lower secondary schools. Education and Information Technologies 13, 221–230 (2008)
7. Jeremic, Z., Jovanovic, J., Gaševic, D.: Evaluating an Intelligent Tutoring System for Design Patterns: the DEPTHS Experience. Journal of Educational Technology & Society 12(2), 111–130 (2009)
8. Kozierkiewicz, A.: Determination of Opening Learning Scenarios in Intelligent Tutoring Systems. In: Zgrzywa, A., Choroś, K., Siemiński, A. (eds.) New Trend in Multimedia and Network Information Systems, pp. 204–213. IOS Press (2008)
9. Kozierkiewicz-Hetmańska, A.: A Conception for Modification of Learning Scenario in an Intelligent E-learning System. In: Nguyen, N.T., Kowalczyk, R., Chen, S.-M. (eds.) ICCCI 2009. LNCS (LNAI), vol. 5796, pp. 87–96. Springer, Heidelberg (2009)
10. Kozierkiewicz-Hetmańska, A., Nguyen, N.T.: A Computer Adaptive Testing Method for Intelligent Tutoring Systems. In: Setchi, R., Jordanov, I., Howlett, R.J., Jain, L.C. (eds.) KES 2010, Part I. LNCS, vol. 6276, pp. 281–289. Springer, Heidelberg (2010)
11. Kozierkiewicz-Hetmańska, A., Nguyen, N.T.: A Method for Scenario Modification in Intelligent E-Learning Systems Using Graph-Based Structure of Knowledge. In: Nguyen, N.T., Katarzyniak, R., Chen, S.-M. (eds.) Advances in Intelligent Information and Database Systems. SCI, vol. 283, pp. 169–179. Springer, Heidelberg (2010)
12. Kozierkiewicz-Hetmańska, A., Nguyen, N.T.: A method for learning scenario determination and modification in intelligent tutoring systems. Applied Mathematics and Computer Science 21(1), 69–82 (2011)
13. Kozierkiewicz-Hetmańska, A.: A Method for Scenario Recommendation in Intelligent E-Learning Systems. Cybernetics and Systems 42(2), 82–99 (2011)
14. Kwok, M., Jones, C.: Catering for different learning styles. Association for Learning Technology Journal (ALT-J) 3(1), 5–11 (1985)
15. Pillay, H., Willss, L.: Computer assisted instruction and individual cognitive style preferences in learning: Does it matter? Australian Educational Computing 11(2), 28–33 (1996)
16. Popescu, E.: Adaptation provisioning with respect to learning styles in a Web-based educational system: an experimental study. Journal of Computer Assisted Learning 26(4), 243–257 (2010)
17. da Silva, G.T., Rosatelli, M.C.: Adaptation in Educational Hypermedia Based on the Classification of the User Profile. In: Ikeda, M., Ashley, K.D., Chan, T.-W. (eds.) ITS 2006. LNCS, vol. 4053, pp. 268–277. Springer, Heidelberg (2006)
18. Weber, G., Peter Brusilovsky, P.: ELM-ART: An Adaptive Versatile System for Web-based Instruction. International Journal of Artificial Intelligence in Education 12, 351–384 (2001)

Knowledge Discovery by an Intelligent Approach Using Complex Fuzzy Sets

Chunshien Li and Feng-Tse Chan

Laboratory of Intelligent Systems and Applications
Department of Information Management, National Central University, Taiwan (R.O.C.)
jamesli@mgt.ncu.edu.tw

Abstract. In the age of rapidly increasing volumes of data, human experts have come to the urgent need to extract useful information from the huge amount of data. Knowldege discovery in databases has obtained much attention for researches and applications in business and in science. In this paper, we present a neuro-fuzzy approach using complex fuzzy sets (CFSs) for the problem of knowledge discovery. A CFS is an advanced fuzzy set, whose membership is complex-valued and characterized by an amplitude function and a phase function. The application of CFSs to the proposed complex neuro-fuzzy system (CNFS) can increase the functional mapping ability to find missing data for knowledge discovery. Moreover, we devise a hybrid learning algorithm to evolve the CNFS for modeling accuracy, combining the artificial bee colony algorithm and the recursive least squares estimator method. The proposed approach to knowledge discovery is tested through experimentation, whose results are compared with those by other approaches. The experimental results indicate that the proposed approach outperforms the compared approaches.

Keywords: complex fuzzy set (CFS), complex neuro-fuzzy system (CNFS), knowledge discovery, artificial bee colony (ABC), recursive least squares estimator (RLSE).

1 Introduction

In the era of the modern information world, data have been accumulating exponentially in various forms. We now have been facing the urgent need to develop new theories and tools to extract useful information from data automatically and effectively. If it is understandable and interpretable by human beings, information can be turned into knowledge. For knowledge discovery and system modeling, several artificial intelligence (AI) based approaches have been presented, where fuzzy logic, neural networks, and neuro-fuzzy systems (NFSs) have been playing significantly important roles for modeling and applications. In general, fuzzy-system-based approaches have the advantage of representing knowledge in the form of If-Then rules, which are transparent to human beings. In contrast, neural-network-based approaches are with the merit of adaptability. The hybrid of fuzzy and neural models has been obtaining the popularity in the community of research, although each of

J.-S. Pan, S.-M. Chen, N.T. Nguyen (Eds.): ACIIDS 2012, Part I, LNAI 7196, pp. 320–329, 2012.

AI-based approaches still has its own important feature in the perspective of research. Zhang et al. [1] used granular neural networks for data fusion and knowledge discovery. Fayyad et al. [2] presented a good overview for knowledge discovery, where the various methods of data mining provide just a single step for the whole process of information discovery. Castellano et al. [3] presented a neuro-fuzzy modeling framework for knowledge discovery. Qin et al. [4] proposed a kernel-based imputation to deal with missing data in knowledge discovery. In [5], Zhang et al. presented a fuzzy modeling approach for training data selection and multi-object optimization. In [6], Rezaee and Zarandi developed a data-driven TSK fuzzy approach for fuzzy modeling. Juang [7] presented the design of recurrent neural network using a hybrid method, which uses genetic algorithm and particle swarm optimization for modeling. Kurban and Beşdok [8] investigated several training methods for a RBF neural network in the application of terrain classification. Boskovitz and Guterman [9] used a neuro-fuzzy system for image segmentation and edge detection. Cpałka [10] designed a neuro-fuzzy classification system. Jang [11] presented the famous adaptive neural-network-based fuzzy inference system (ANFIS), which molds the hybrid of a neural network and a fuzzy inference system, for system modeling and forecasting. Scherer in [12] used a neuro-fuzzy system for nonlinear modeling. In general, neural fuzzy systems [13] are excellent tools for modeling and for knowledge discovery. Qin and Yang [14] studied a neuro-fuzzy method for image noise cancellation. In [15], Zounemat-Kermani and Teshnehlab presented a neuro-fuzzy approach to time series forecasting.

In this paper, we present a framework called the complex neuro-fuzzy system (CNFS) for knowledge discovery. A CNFS is a neuro-fuzzy based system whose kernel is composed of fuzzy If-Then rules which are characterized by complex fuzzy sets (CFSs). Ramot at al. [16] proposed the theory of CFS. A CFS is an advanced fuzzy set whose membership degree is complex-valued and characterized by an amplitude function and a phase function in the unit disc of the complex plane. The property of complex membership state of a CFS indeed makes difference from a traditional type-1 fuzzy set, whose membership degree is defined normally in the real-valued unit interval of [0, 1]. The application of CFSs [17]-[19] to the design of adaptive systems can increase their adaptability for learning, so that the functional mapping capability of the adaptive systems can be augmented. For this motivation, we proposed the CNFS approach using CFSs for the process of knowledge discovery. For knowledge discovery, the goal of the study is that with the proposed CNFS approach we try to represent numerical-linguistic records [1] by rules. In the perspective of If-Then rules, we apply the proposed CNFS models to create a framework of knowledge discovery, which can transform these numerical-linguistic records into fuzzy If-Then rules. And, in the principle of divide-and-conquer, the parameters of CNFS are imaginatively separated into two groups: the premise parameters and the consequent parameters. For the parameter learning of CNFS, we devise a hybrid ABC-RLSE learning method, which integrates the famous artificial bee colony (ABC) algorithm [20]-[22] and the well-known recursive least squares estimator (RLSE) method [13]. The ABC-RLSE algorithm is applied to evolve the parameters of CNFS, in the way that the ABC is used to update the premise parameters and the RLSE is used to update the consequent parameters. The ABC-RLSE can achieve fast training for the proposed approach to the application of knowledge discovery.

We arrange the rest of the paper in the following. In Section 2, we present the knowledge discovery framework. In Section 3, the proposed CNFS and the ABC-RLSE learning method for knowledge discovery are specified. In Section 4, Experimentation is conducted to test the proposed approach for knowledge discovery. The experimental results are compared to other approaches. Finally, the paper is concluded.

2 Framework for Knowledge Discovery

In this paper, for knowledge discovery, we adopt and improve the framework in [1]. The improved framework has three parts, including the feature extraction for numerical-linguistic records from a dataset, the CNFS modeling with the extracted records, and the linguistic records output. The framework is shown in Fig. 1.

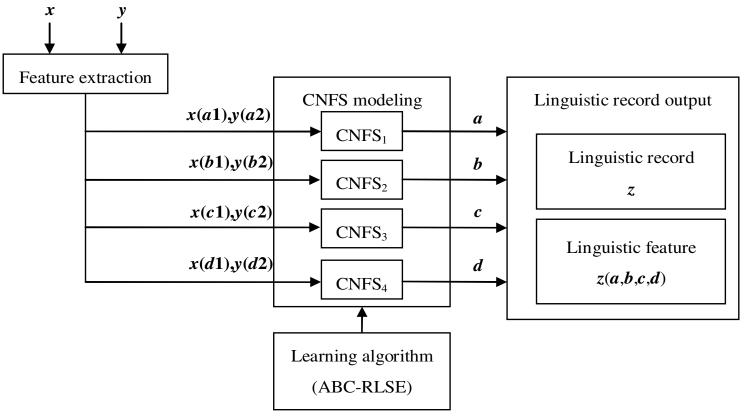


Fig. 1. CNFS-based framework for knowledge discovery

(A) ***Feature Extraction*****.** The block of feature extraction for numerical-linguistic records, shown in Fig. 1, is used to extract feature for each datum. By the concept of trapezoidal feature vector in [1], we can have a feature vector (w, $w/10$, $w/2$, $w/2$) for a datum w. For example, in Fig. 1, there are two inputs, x and y, to the block of feature extraction. For the first input, x, the trapezoidal feature vector is denoted as ($x(a1)$, $x(b1)$, $x(c1)$, $x(d1)$)= (x, $x/10$, $x/2$, $x/2$). Similarly, the feature vector for the second input, y, is denoted as ($y(a2)$, $y(b2)$, $y(c2)$, $y(d2)$)= (y, $y/10$, $y/2$, $y/2$).

(B) ***CNFS Modeling*****.** As shown in Fig. 1, the block of CNFS modeling consists of four CNFS models, each of which is composed of a set of fuzzy If-Then rules. For the first CNFS (denoted as CNFS_1), the first components of the feature vectors for x and y, which are the $x(a1)$ and the $y(a2)$, are used as two inputs to CNFS_1, whose output is then used as the first component of the feature vector for z. Similar operations can be applied to the others {CNFS_i, i=2,3,4}. The goal for the block of CNFS modeling is to produce a feature vector for z, which is the output of the system for knowledge discovery. To optimize these CNFS models, the ABC-RLSE hybrid learning method is applied. The optimization is based on the measure of root mean square error (RMSE). Note that the details for theory of CNFS and the ABC-RLSE learning method are specified later in the following section.

(C) *Linguistic Record Output.* The block of linguistic record output has two subparts. One is to generate the linguistic record *z*, and the other is to produce the feature vector, denoted as (*a*,*b*,*c*,*d*), for *z*.

3 Methodology

Based on the concept of CFS, the membership function of a CFS is complex-valued and characterized within the unit disc of the complex plan. The general form for the complex-valued membership function of a CFS is given below.

$$\begin{aligned} \mu_A(h) &= r_A(h)\exp\big(j\omega_A(h)\big) \\ &= \mathrm{Re}\big(\mu_A(h)\big) + j\mathrm{Im}\big(\mu_A(h)\big) \\ &= r_A(h)\cos\big(\omega_A(h)\big) + jr_{\mathrm{A}}(h)\sin\big(\omega_A(h)\big), \end{aligned} \quad (1)$$

where $\mu_A(h)$ is the membership function; $j = \sqrt{-1}$; A denotes the CFS; h is the input base variable; $r_A(h)$is the amplitude function; $\omega_A(h)$is the phase function. We devise a Gaussian complex fuzzy set (GCFS) for the design of the premises of CNFS. The general form of a GCFS is characterized below.

$$\mathrm{GCFS}(h, m, \sigma) = \exp\left[-\frac{(h-m)^2}{2\sigma^2}\right] + j\frac{-(h-m)}{\sigma^2}\exp\left[-\frac{(h-m)^2}{2\sigma^2}\right], \quad (2)$$

where $j = \sqrt{-1}$; m and σ are the mean and the spread of the GCFS, respectively.

Suppose that we have a CNFS that consists of *K* first-order Takagi-Sugeno (T-S) fuzzy rules, given as follows.

$$\begin{aligned} &\text{Rule } i: \text{ IF } \left(x_1 \text{ is } A_1^{(i)}(h_1)\right) \text{ and } \left(x_2 \text{ is } A_2^{(i)}(h_2)\right) \cdots \text{and } \left(x_M \text{ is } A_M^{(i)}(h_M)\right) \\ &\qquad \text{Then } z^{(i)} = a_0^{(i)} + \textstyle\sum_{j=1}^{M} a_j^{(i)} h_j, \end{aligned} \quad (3)$$

for *i* =1,2,...,*K*, where *i* is the index for the *i*th fuzzy rule; *M* is the number of inputs; x_j is the *j*th input linguistic variable for *j*=1,2,…,*M*; h_j is the *j*th input base variable; $A_j^{(i)}(h_j)$ is the *j*th premise, which is defined by a GCFS in (2); $z^{(i)}$ is the nominal output; {$a_j^{(i)}$, *j*=0,1,…,*M*} are consequent parameters. The complex fuzzy inference process can be cast into a neural-net structure to become a CNFS with six layers, specified below.

Layer 0: This layer receives the inputs and transmits them to Layer 1. The input vector at time *t* is given as follows.

$$H(t) = [h_1(\mathrm{t})\, h_2(\mathrm{t}) \cdots h_M(t)]^{\mathrm{T}}. \tag{4}$$

Layer 1: This layer is called the complex fuzzy-set layer, which is used to calculate complex membership degrees of GCFSs. **Layer 2:** This layer is called the firing-strength layer. The firing strength of the *i*th rule is calculated and defined as fellow.

$$\beta^{(i)}(t) = \prod_{j=1}^{M} r_j^{(i)}\big(h_j(t)\big) \exp\Big(j\omega_{A_1^{(i)} \cap \ldots \cap A_M^{(i)}}\Big), \tag{5}$$

where $r_j^{(i)}$ and $\omega_{A_1^{(i)} \cap \ldots \cap A_M^{(i)}}$ are the amplitude function and the phase function of the firing strength of the *i*th rule. **Layer 3:** This layer is used for the normalization of the firing strengths. The normalized firing strength for the *i*th rule is given as follows.

$$\lambda^{(i)}(t) = \frac{\beta^{(i)}(t)}{\sum_{i=1}^{K} \beta^{(i)}(t)} = \frac{\prod_{j=1}^{M} r_j^{(i)}(h_j(t)) \exp\Big(j\omega_{A_1^{(i)} \cap \ldots \cap A_M^{(i)}}\Big)}{\sum_{i=1}^{K} \prod_{j=1}^{M} r_j^{(i)}(h_j(t)) \exp\Big(j\omega_{A_1^{(i)} \cap \ldots \cap A_M^{(i)}}\Big)}. \tag{6}$$

Layer 4: The layer is called the consequent layer. The normalized consequent of the *i*th rule is represented as follows.

$$\begin{aligned} \xi^{(i)}(t) &= \lambda^{(i)}(t)\ z^{(i)}(t) = \lambda^{(i)}(t)\big(a_0^{(i)} + \textstyle\sum_{j=1}^{M} a_j^{(i)} h_j(t)\big), \\ &= \frac{\prod_{j=1}^{M} r_j^{(i)}(h_j(t)) \exp\Big(j\omega_{A_1^{(i)} \cap \ldots \cap A_M^{(i)}}\Big)}{\sum_{i=1}^{K} \prod_{j=1}^{M} r_j^{(i)}(h_j(t)) \exp\Big(j\omega_{A_1^{(i)} \cap \ldots \cap A_M^{(i)}}\Big)} \Big(a_0^{(i)} + \textstyle\sum_{j=1}^{M} a_j^{(i)} h_j(t)\Big). \end{aligned} \tag{7}$$

Layer 5: This layer is called the output layer. The normalized consequents from Layer 4 are congregated into the layer to produce the CNFS output, given as follows.

$$\xi(t) = \sum_{i=1}^{K} \xi^{(i)}(t) = \sum_{i=1}^{K} \lambda^{(i)}(t)\ z^{(i)}(t). \tag{8}$$

The complex-valued output of the CNFS can be expressed as follows.

$$\xi(t) = \xi_{\mathrm{Re}}(t) + j\xi_{\mathrm{Im}}(t) = |\xi(t)|\cos\big(j\omega_\xi\big) + j|\xi(t)|\sin\big(j\omega_\xi\big), \tag{9}$$

where $\xi_{\mathrm{Re}}(t)$ is the real part of the output, and $\xi_{\mathrm{Im}}\ (t)$ is the imaginary part. Based on (8), the CNFS can be viewed as a complex-valued function, expressed as follows.

$$\xi(t) = \mathrm{F}(\mathbf{H}(t), \mathbf{W}) = \mathrm{F}_{\mathrm{Re}}(\mathbf{H}(t), \mathbf{W}) + j\mathrm{F}_{\mathrm{Im}}(\mathbf{H}(t), \mathbf{W}), \tag{10}$$

where $\mathrm{F}_{\mathrm{Re}}(.)$ is the real part of the CNFS output; $\mathrm{F}_{\mathrm{Im}}(.)$ is the imaginary part; $\mathbf{H}(t)$ is the input vector to the CNFS; **W** denotes the parameter set of the CNFS, including the premise parameters and the consequent parameters, denoted as $\mathbf{W}_{\mathrm{If}}$ and $\mathbf{W}_{\mathrm{Then}}$, respectively.

$$\mathbf{W} = \mathbf{W}_{\mathrm{If}} \cup \mathbf{W}_{\mathrm{Then}}. \tag{11}$$

For the training of the proposed CNFS, we apply the ABC-RLSE hybrid learning method, where $\mathbf{W}_{\text{If}}$ and $\mathbf{W}_{\text{Then}}$ are updated by the ABC and RLSE, respectively, in a hybrid way.

The artificial bee colony (ABC) algorithm is an optimization method which simulates the nectar-searching behavior by bees. Basically, the bees of a bee colony can be separated into three groups: the employed bees (EBs), the onlooker bees (OBs), and the scout bees (SBs). The EBs perform their job by flying to the food sources and bringing back food to the colony. Moreover, these EBs reveal the locations of food sources to the OBs by dancing. With the dancing information, each of the OBs selects a food source to go. The selection of a food source is dependent on the nectar amount of each food source. For the SBs, they fly randomly to look for new food sources. Note that the nectar amount of each food source corresponds to a fitness value for a specific optimization problem and the food source location represents a candidate solution to the optimization problem. For the ABC algorithm, assume there are S food sources. The location of the ith food source is expressed as $\mathbf{X}_i$=[x_{i1}, x_{i2}, …, x_{iQ}] for i=1,2,…,S. In the ABC algorithm, the location of the ith food source is updated by the following equation.

$$x_{ij}(t+1) = x_{ij}(t) + \varphi_{ij}\left(x_{ij}(t) - x_{kj}(t)\right), \tag{12}$$

for i=1,2,…,S and j=1,2,…,Q, where $x_{ij}(t)$ is the jth dimensional coordinate of the ith food source at iteration t, k is a random integer in the set of {1,2,…,S} with the constraint of i≠k, and φ_{ij} is a value between [-1, 1]. An onlooker bee goes to the vicinity of X$_i$ with the probability given below.

$$p_i = \frac{F_{fitness}(\mathbf{X}_i)}{\sum_{j=1}^{S} F_{fitness}(\mathbf{X}_j)}, \tag{13}$$

where $F_{fitness}(.)$ is the fitness function. In the ABC, if the fitness of a food source is not improved further through a predetermined number of cycles, called *limit*, then that food source is assumed to be abandoned. The operation of the ABC is specified in steps. **Step 1:** initialize the locations of bees. **Step 2:** send employed bees to food sources and compute their fitness values. **Step 3:** send onlooker bees to the food sources, according to (13), and then compute their fitness values. **Step 4:** send scout bees for other food sources to help find better food source as possible. **Step 5:** update the locations of food sources and save the best one so far. **Step 6:** if any termination condition is met, stop; otherwise increase the iteration index and go back to **Step 2** to continue the procedure.

The recursive least squares estimator (RLSE) is a recursive method for the problem of least squares estimation (LSE). The model of LSE is given below.

$$y = \theta_1 f_1(u) + \theta_2 f_2(u) + \cdots + \theta_m f_m(u) + \varepsilon, \tag{14}$$

where y is the target; u is the input to model; {$f_i(u)$, i=1,2,..,m} are known functions of u; {θ_i , i=1,2,…,m} are unknown parameters to be estimated; ε is the model error. Note that the parameters {θ_i, i=1,2,…,m} can be viewed as the consequent parameters of the proposed CNFS. The observed samples are collected to be used as training data for the proposed CNFS. The training data (TD) is denoted as follows.

$$\mathrm{TD} = \{(u_i, y_i), i = 1,2, \dots, N\}, \tag{15}$$

where (u_i, y_i) is the ith data pair in the form of (*input*, *target*). With (15), the LSE model can be expressed in matrix form, $\mathbf{y}=\mathbf{A}\boldsymbol{\theta}+\boldsymbol{\varepsilon}$, where $\mathbf{y} = [y_1\, y_2 \dots y_N]^{\mathrm{T}}$; $\boldsymbol{\theta} = [\theta_1\, \theta_2 \dots \theta_m]^{\mathrm{T}}$; $\boldsymbol{\varepsilon} = [\varepsilon_1\, \varepsilon_2 \dots \varepsilon_N]^{\mathrm{T}}$; **A** is the matrix which is formed by the known functions $\{f_i(u_j), i=1,2,..,m\}$ for $j=1,2,\dots,N$.

The optimal estimation for $\boldsymbol{\theta}$ can be calculated using the following RLSE equations recursively.

$$\mathbf{P}_{k+1} = \mathbf{P}_k - \frac{\mathbf{P}_k \mathbf{b}_{k+1} \mathbf{b}_{k+1}^T \mathbf{P}_k}{1+\mathbf{b}_{k+1}^T \mathbf{P}_k \mathbf{b}_{k+1}}, \tag{16}$$

$$\boldsymbol{\theta}_{k+1} = \boldsymbol{\theta}_k + \mathbf{P}_{k+1}\mathbf{b}_{k+1}(y_{k+1} - \mathbf{b}_{k+1}^T \boldsymbol{\theta}_k), \tag{17}$$

for $k = 0,1,2,\dots,(N\text{-}1)$, where $\boldsymbol{\theta}_k$ is the estimator at the kth iteration; $[\mathbf{b}_{k+1}^{\mathrm{T}}, y_{k+1}]$ is the (k+1)th row of [**A, y**]. To implement the RLSE, we initialize $\boldsymbol{\theta}_0$ with the zero vector and $\mathbf{P}_0=\alpha\mathbf{I}$, where α is a large positive number and **I** is the identity matrix.

4 Experimentation

The target of this experiment is to find the unknown records and represents the unknown records by *IF-Then* rules. We use the trapezoidal feature vector to represent a linguistic record [1]. Suppose we have a data set for (x, y, z) that includes numerical and linguistic data, as shown in Table 1, in which x and y are the two variables for inputs; z is the variable for output linguistic record; $\tilde{d}$ denotes a input linguistic record, meaning it is around the value d; "?" denotes that the output linguistic record is missing. For the linguistic record of x, there is a feature vector by ($x(a1)$, $x(b1)$, $x(c1)$, $x(d1)$). Similarly, for the linguistic record of y, there is a feature vector by ($y(a2)$, $y(b2)$, $y(c2)$, $y(d2)$). For the record of z, a multiplication $z= xy$ is implied, whose feature vector for linguistic representation is represented by (a, b, c, d). There are nine fuzzy If-Then rules to represent these known records, shown in the Table 2.

Table 1. Known and unknown numerical-linguistic records

		y				
		$\tilde{1}$	$\widetilde{1.5}$	$\tilde{2}$	2.5	$\tilde{3}$
x	$\tilde{1}$	$\tilde{1}$	?	$\tilde{2}$	?	$\tilde{3}$
	1.5	?	?	?	?	?
	$\tilde{2}$	$\tilde{2}$	?	$\tilde{4}$	?	$\tilde{6}$
	$\widetilde{2.5}$	?	?	?	?	?
	$\tilde{3}$	$\tilde{3}$	?	$\tilde{6}$	?	$\tilde{9}$

$\tilde{d}$: linguistic record around a numerical value d.

? : unknown record.

Table 2. Nine fuzzy rules of known linguistic records

x, (*x*(*a*1), *x*(*b*1), *x*(*c*1), *x*(*d*1))	*y*, (*y*(*a*2), *y*(*b*2), *y*(*c*2), *y*(*d*2))	*z* = *xy*, (*a*, *b*, *c*, *d*)
around 1, (1, 0.1, 0.5, 0.5)	*around* 1, (1, 0.1, 0.5, 0.5)	*around* 1, (1, 0.1, 0.5, 0.5)
around 1, (1, 0.1, 0.5, 0.5)	*around* 2, (2, 0.2, 1.0, 1.0)	*around* 2, (2, 0.2, 1.0, 1.0)
around 1, (1, 0.1, 0.5, 0.5)	*around* 3, (3, 0.3, 1.5, 1.5)	*around* 3, (3, 0.3, 1.5, 1.5)
around 2, (2, 0.2, 1.0, 1.0)	*around* 1, (1, 0.1, 0.5, 0.5)	*around* 2, (2, 0.2, 1.0, 1.0)
around 2, (2, 0.2, 1.0, 1.0)	*around* 2, (2, 0.2, 1.0, 1.0)	*around* 4, (4, 0.4, 2.0, 2.0)
around 2, (2, 0.2, 1.0, 1.0)	*around* 3, (3, 0.3, 1.5, 1.5)	*around* 6, (6, 0.6, 3.0, 3.0)
around 3, (3, 0.3, 1.5, 1.5)	*around* 1, (1, 0.1, 0.5, 0.5)	*around* 3, (3, 0.3, 1.5, 1.5)
around 3, (3, 0.3, 1.5, 1.5)	*around* 2, (2, 0.2, 1.0, 1.0)	*around* 6, (6, 0.6, 3.0, 3.0)
around 3, (3, 0.3, 1.5, 1.5)	*around* 3, (3, 0.3, 1.5, 1.5)	*around* 9, (9, 0.9, 4.5, 4.5)

In order to discover the 16 unknown linguistic records for *z* in Table 1, we use four CNFS models to discover their feature vectors. The framework of knowledge discovery is shown in Fig. 1. The feature vectors of the known (*x*, *y*, *z*) linguistic record pairs are used as the training data for the four CNFS models, for which the ABC-RLSE method is applied for the training purpose. The settings for the ABC-RLSE method are shown in Table 3.

Table 3. Settings of the ABC-RLSE method

Settings of ABC		Settings of RLSE	
Bee colony dimensions	20	Number of consequent parameters	75
Bee colony size	4	θ	75×1
Scout bee	1	A	10^8
Maximal iterations	200	I (identity matrix)	75×1

After learning, the CNFS models can generate feature vectors for the sixteen unknown linguistic records, respectively, as shown in Table 4. The discovered linguistic records are represented by *If-Then* rules, as shown in Table 5. The results are compared to other approaches. Performance comparison in the measure of average absolute error (ABE) [1] is given in Table 6.

Table 4. New linguistic records discovered by the proposed approach

	$\tilde{1}$	$\widetilde{1.5}$	$\tilde{2}$	2.5	$\tilde{3}$
$\tilde{1}$	$\tilde{1}$	$1.\widetilde{4940}$	$\tilde{2}$	$2.\widetilde{5067}$	$\tilde{3}$
1.5	$1.\widetilde{4888}$	$2.\widetilde{2241}$	$2.\widetilde{9770}$	$3.\widetilde{7311}$	$4.\widetilde{4649}$
$\tilde{2}$	$\tilde{2}$	$2.\widetilde{9880}$	$\tilde{4}$	$5.\widetilde{0136}$	$\tilde{6}$
$\widetilde{2.5}$	$2.\widetilde{5114}$	$3.\widetilde{7523}$	$5.\widetilde{0235}$	$6.\widetilde{2969}$	$7.\widetilde{5359}$
$\tilde{3}$	$\tilde{3}$	$4.\widetilde{4818}$	$\tilde{6}$	$7.\widetilde{5207}$	$\tilde{9}$

Table 5. New 16 rules discovered by the proposed approach

x	*y*	*z*, (*a*, *b*, *c*, *d*)
around 1	*around* 1.5	*around* 1.4940, (1.4940, 0.1499, 0.7498, 0.7542)
around 1	2.5	*around* 2.5067, (2.5067, 0.2500, 1.2501, 1.2409)
1.5	*around* 1	*around* 1.4888, (1.4888, 0.1499, 0.7489, 0.7482)
1.5	*around* 1.5	*around* 2.2241, (2.2241, 0.2249, 1.1231, 1.1288)
1.5	*around* 2	*around* 2.9970, (2.9970, 0.3000, 1.4978, 1.4965)
1.5	2.5	*around* 3.7311, (3.7311, 0.3750, 1.8724, 1.8570)
1.5	*around* 3	*around* 4.4649, (4.4649, 0.4500, 2.2467, 2.2448)
around 2	*around* 1.5	*around* 2.9880, (2.9880, 0.2999, 1.4997, 1.5085)
around 2	2.5	*around* 5.0136, (5.0136, 0.5000, 2.5002, 2.4819)
around 2.5	*around* 1	*around* 2.5114, (2.5114, 0.2500, 1.2511, 1.2521)
around 2.5	*around* 1.5	*around* 3.7523, (3.7523, 0.3750, 1.8764, 1.8888)
around 2.5	*around* 2	*around* 5.0235, (5.0235, 0.4999, 2.5023, 2.5043)
around 2.5	2.5	*around* 6.2969, (6.2969, 0.6249, 3.1282, 3.1079)
around 2.5	*around* 3	*around* 7.5359, (7.5359, 0.7499, 3.7535, 3.7564)
around 3	*around* 1.5	*around* 4.4818, (4.4818, 0.4500, 2.2496, 2.2626)
around 3	2.5	*around* 7.5207, (7.5207, 0.7499, 3.7503, 3.7231)

Table 6. Performance comparison

	ABE	
Method	Training phase	Testing phase
CGNN [1]	0.0001	0.303
FGNN [1]	0.0001	0.224
Proposed	5.58×10^{-8}	0.0194

5 Conclusion

The complex neuro-fuzzy system (CNFS) based approach using complex fuzzy sets for knowledge discovery has been presented. The ABC-RLSE hybrid learning method has been devised for training the proposed CNFS models in the framework of knowledge discovery. Through experimentation, the proposed approach has shown excellent performance in finding missing data for knowledge discovery. The experimental results indicate that our proposed approach outperforms the compared methods in performance comparison.

References

1. Zhang, Y.Q., Fraser, M.D., Gagliano, R.A., Kandel, A.: Granular neural networks for numerical-linguistic data fusion and knowledge discovery. IEEE Transactions on Neural Networks 11, 658–667 (2000)
2. Fayyad, U., Piatetsky-Shapiro, G., Smyth, P.: The KDD process for extracting useful knowledge from volumes of data. Commun. ACM 39(11), 27–34 (1996)

3. Castellano, G., Castiello, C., Fanelli, A.M., Mencar, C.: Knowledge discovery by a neuro-fuzzy modeling framework. Fuzzy Sets and Systems 149, 187–207 (2005)
4. Qin, Y., Zhang, S., Zhu, X., Zhang, J., Zhang, C.: POP algorithm: Kernel-based imputation to treat missing values in knowledge discovery from databases. Expert Systems with Applications 36, 2794–2804 (2009)
5. Zhang, Q., Mahfouf, M.: A hierarchical Mamdani-type fuzzy modelling approach with new training data selection and multi-objective optimisation mechanisms: A special application for the prediction of mechanical properties of alloy steels. Applied Soft Computing 11, 2419–2443 (2011)
6. Rezaee, B., Zarandi, M.H.F.: Data-driven fuzzy modeling for Takagi–Sugeno–Kang fuzzy system. Information Sciences 180, 241–255 (2010)
7. Juang, C.F.: A hybrid of genetic algorithm and particle swarm optimization for recurrent network design. IEEE Transactions on Systems, Man, and Cybernetics 34, 997–1006 (2004)
8. Kurban, T., Beşdok, E.: A comparison of RBF neural network training algorithms for inertial sensor based terrain classification. Sensors, 6312–6329 (2009)
9. Boskovitz, V., Guterman, H.: An adaptive neuro-fuzzy system for automatic image segmentation and edge detection. IEEE Transactions on Fuzzy Systems 10, 247–262 (2002)
10. Cpałka, K.: A new method for design and reduction of neuro-fuzzy classification systems. IEEE Transactions on Neural Networks 20, 701–714 (2009)
11. Jang, S.R.: ANFIS: adaptive-network-based fuzzy inference system. IEEE Transactions on Systems, Man, and Cybernetics 23, 665–685 (1993)
12. Scherer, R.: Neuro-fuzzy relational systems for nonlinear approximation and prediction. Nonlinear Analysis: Theory, Methods & Applications 71, 1420–1425 (2009)
13. Jang, J.S.R., Sum, C.T., Mizutani, E.: Neuro-fuzzy and soft computing. Prentice-Hall, Englewood Cliffs (1997)
14. Qin, H., Yang, S.X.: Adaptive neuro-fuzzy inference systems based approach to nonlinear noise cancellation for images. Fuzzy Sets and Systems 158, 1036–1063 (2007)
15. Zounemat-Kermani, M., Teshnehlab, M.: Using adaptive neuro-fuzzy inference system for hydrological time series prediction. Applied Soft Computing 8, 928–936 (2008)
16. Ramot, D., Milo, R., Friedman, M., Kandel, A.: Complex fuzzy sets. IEEE Transactions on Fuzzy Systems 10, 171–186 (2002)
17. Chen, Z., Aghakhani, S., Man, J., Dick, S.: ANCFIS: A neurofuzzy architecture employing complex fuzzy sets. IEEE Transactions on Fuzzy Systems 19, 305–322 (2011)
18. Dick, S.: Toward complex fuzzy logic. IEEE Transactions on Fuzzy Systems 13, 405–414 (2005)
19. Aghakhani, S., Dick, S.: An on-line learning algorithm for complex fuzzy logic. In: IEEE International Conference on Fuzzy Systems (FUZZ), pp. 1–7 (2010)
20. Irani, R., Nasimi, R.: Application of artificial bee colony-based neural network in bottom hole pressure prediction in underbalanced drilling. J. Petroleum Science and Engineering 78, 6–12 (2011)
21. Karaboga, D., Basturk, B.: A powerful and efficient algorithm for numerical function optimization: artificial bee colony (ABC) algorithm. J. Global Optimization 39, 171–459 (2007)
22. Ozturk, C., Karaboga, D.: Hybrid artificial bee colony algorithm for neural network training. In: IEEE Congress on Evolutionary Computation (CEC), pp. 84–88 (2011)

Integration of Multiple Fuzzy FP-trees

Tzung-Pei Hong[1,3], Chun-Wei Lin[1,*], Tsung-Ching Lin[1],
Yi-Fan Chen[2], and Shing-Tai Pan[1]

[1] Department of Computer Science and Information Engineering
[2] Department of Applied Mathematics
National University of Kaohsiung
Kaohsiung, Taiwan, R.O.C.
[3] Department of Computer Science and Engineering
National Sun Yat-sen University
Kaohsiung, Taiwan, R.O.C.
{tphong,stpan}@nuk.edu.tw, jerrylin@ieee.org,
{m0985507,A0974154}@mail.nuk.edu.tw

Abstract. In the past, the MFFP-tree algorithm was proposed to handle the quantitative database for efficiently mining the complete fuzzy frequent itemsets. In this paper, we propose an integrated MFFP (called iMFFP)-tree algorithm for merging several individual MFFP trees into an integrated one. It can help derive global fuzzy rules among distributed databases, thus allowing managers to make more sophisticated decisions. Experimental results also showed the performance of the proposed approach.

Keywords: iMFFP tree, integration, fuzzy data mining, quantitative database, distributed database.

1 Introduction

Depending on the variety of knowledge desired, data mining approaches can be divided into association rules [1, 3], classification [10, 18], clustering [12, 13], and sequential patterns [2, 17], among others. Among them, mining association rules from databases is especially common in data mining research [5, 6, 16]. Many algorithms were used for processing the whole database to find the desired information. In real-world applications, however, a parent company may own multiple branches, and each branch has its own local database. The manager in a parent company needs to make a decision for the entire company from the collected databases in different branches. Thus, it is important to efficiently integrate many different databases to make a usable decision.

In the past, Lan and Qiu proposed a novel parallel algorithm called PFPT algorithm [14]. It merged several FP trees into an integrated one, thus avoiding the FP-growth mining approach to generate a huge number of intermediate results. In this paper, we extend the PFPT algorithm and propose a MFFP-tree merging algorithm for

[*] Corresponding author.

J.-S. Pan, S.-M. Chen, N.T. Nguyen (Eds.): ACIIDS 2012, Part I, LNAI 7196, pp. 330–337, 2012.

integrating different databases into one, forming an integrated MFFP (called iMFFP) tree. The iMFFP tree inherits the property of a MFFP tree for handing quantitative databases in fuzzy data mining.

The remainder of this paper is organized as follows. The related work is described in section 2. The proposed algorithm for integrating multiple MFFP trees is stated in section 3. An example to illustrate the propsoed algorithm is mentioned in section 4. Experimental results are given in section 5. Conclusions are provided in section 6.

2 Review of Related Work

In this section, some related researches are briefly reviewed. They consist of the fuzzy data mining approaches and the frequent pattern tree.

2.1 Fuzzy Data Mining Approaches

In recent years, fuzzy set theory [19] has been used frequently in intelligent systems because of its simplicity and similarity to human reasoning. Kuok *et al.* proposed a fuzzy mining approach for handling numerical data in databases and deriving fuzzy association rules [11]. Hong *et al.* proposed fuzzy mining algorithms for discovering fuzzy rules from quantitative transaction data [8]. Papadimitriou *et al.* proposed an approach based on FP-trees for finding fuzzy association rules [15]. In addition, Hong *et al.* also proposed the multiple fuzzy frequent pattern tree algorithm [9] to efficiently mine complete fuzzy frequent itemsets from quantitative databases, which extended the FP-tree mining process for constructing its suitable tree structures.

2.2 The Frequent Pattern Trees

Frequent pattern mining is one of the most important research issues in data mining. The initial algorithm for mining association rules was given by Agrawal *et al.* [1] in the form of Apriori algorithm which is based on level-wise candidate set generation and test methodology. However, because the size of the database can be very large, it is very costly to repeatedly scan the database to count supports for candidate itemsets. The limitations of the Apriori algorithm are overcome by an innovative approach; the frequent pattern (FP) tree structure and the FP-growth algorithm by Han *et al.* [7]. Their approach can efficiently mine frequent itemsets without the generation of candidate itemsets, and it scans the original transaction database only twice. The mining algorithm consists of two phases; the first constructs an FP-tree structure and the second recursively mines the frequent itemsets from the structure. After the FP-growth algorithm is then executed, the frequent itemsets with a given support would be derived from the FP-tree structure.

3 The Proposed Integrated Multiple Fuzzy FP-tree Algorithm

In this section, details of the proposed MFFP-tree merging algorithm are described below.

***The integrated MFFP-tree algorithm*:**

INPUT: Multiple quantitative databases DB_k, each of them consisting of n transactions, a set of membership functions, and a predefined minimum support threshold s.

OUTPUT: An integrated MFFP (iMFFP) tree.

STEP 1: Transform quantitative value v_{ij} of each item I_j in the i-th transaction of each database DB_k into a fuzzy set f_{ij} represented as $(f_{ij1}/R_{j1} + f_{ij2}/R_{j2} + \ldots + f_{ijh}/R_{jh})$ using the given membership functions, where h is the number of fuzzy regions for I_j, R_{jl} is the l-th fuzzy region of I_j, $1 \le l \le h$, and f_{ijl} is v_{ij}'s fuzzy membership value in region R_{jl}. Note that f_{ijl}/R_{jl} means that the membership value of region R_{jl} is f_{ijl}.

STEP 2: Calculate the scalar cardinality $count_{jl}$ of each fuzzy region R_{jl} in the transactions for all DB_k as:

$$count_{j_l} = \sum_{\forall k} \sum_{t=1}^{n} f_{ijl}.$$

STEP 3: Check whether the value $count_{jl}$ of the fuzzy region R_{jl} is larger than or equal to the predefined minimum count $n \times s$. If the count of a fuzzy region R_{jl} is equal to or greater than minimum count, it can be treated as a fuzzy frequent itemset and put it in the set of L_1. That is:

$$L_1 = \{R_{jl} \mid count_{jl} > n \times s,\ 1 \le j \le m\}.$$

STEP 4: Build the sub-MFFP tree of each DB_k $(1 \le k \le N)$, which only keeps the fuzzy regions existing in L_1.

STEP 5: Find all leaf nodes of sub-MFFP tree of DB_k initially for integrating the extracted branches into sub-MFFP tree of DB_{k+1} as an integrated MFFP tree. Trace all leaf nodes in the path from bottom to top to extract the node R_{jl} with the satisfied count forming a branch until it reaches the root node of sub-MFFPk tree, two cases may exist:

Substep 5-1: If the parent-node at level (t-1) of the currently processed node R_{jl} has only one child-node at level t of the currently processed node R_{jl}, node R_{jl} with its count at level (t-1) is then directly extracted to form a pre-integrated node in the branch.

Substep 5-2: Otherwise, set the count of the currently processed node R_{jl} at level (t-1) the same as its child-node at level t in the branch. Extract it to form a pre-integrated node in the branch.

STEP 6: Merge the extracted branches in STEP 5 to form an integrated MFFP (iMFFP) tree. The following two cases may exist.

Substep 6-1: If a fuzzy region R_{jl} in a transaction is at the corresponding branch of iMFFP tree, add the fuzzy value f_{ijl} of R_{jl} in the processed transaction to the node of R_{jl} in the branch.

Substep 6-2: Otherwise, add a node of R_{jl} at the end of the corresponding branch, set the count of the node as the fuzzy value f_{ijl} of R_{jl}, and connect the node of R_{jl} in the last branch with the current node as a sequence.

STEP 7: Repeat the STEPs 4 to 6 until all DB_k have be integrated into one iMFFP tree.

STEP 8: Build an indexed Header_Table by keeping fuzzy frequent itemsets in STEP 3. Insert a node-link from the entry of R_{jl} in the Header_Table to the first branch of node R_{jl}.

4 An Illustrative Example

Assume that there are two quantitative databases DB_1 and DB_2 which shown in Table 1, and the minimum support threshold s is set to 30%. Both of each consisted of 4 transactions and 5 items, denoted {A} to {E}.

Table 1. Two quantitative databases

TID	*Item*	*Database*
1	(A:5) (B:2) (C:5)	DB_1
2	(A:3) (C:10) (D:2) (E:2)	DB_1
3	(A:5) (B:2) (C:8) (E:6)	DB_1
4	(C:9) (D:3)	DB_1
5	(A:5) (C:10) (D:2) (E:9)	DB_2
6	(A:8) (B:2) (C:3)	DB_2
7	(B:3) (C:9)	DB_2
8	(A:7) (C:9) (D:3)	DB_2

Assume that the fuzzy membership functions are the same for all items shown in Figure 1. In this example, amounts are represented by three fuzzy regions: {*Low*}, {*Middle*}, and {*High*}. Thus, three fuzzy membership values are produced for each item in a transaction according to the predefined membership functions.

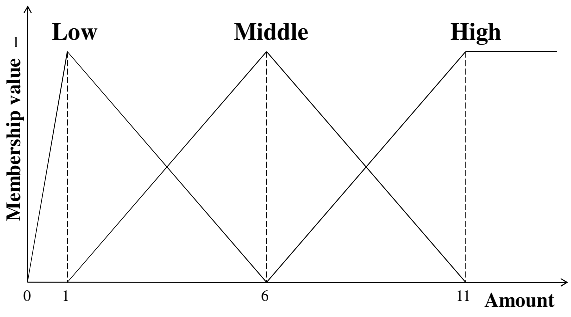


Fig. 1. The used membership functions

STEPs 1 to 3: The quantitative values of the items in the transactions are represented as fuzzy sets using the membership functions. The scalar cardinality of each fuzzy region in the transactions of two databases is calculated as the count value and be checked against the specified minimum count, which is (8×0.3) (= 2.4) to find fuzzy frequent 1-itemsets. The results are shown in Table 2.

Table 2. Counts of fuzzy regions

Fuzzy region	*Count*
A.Middle	4.2
B.Low	3.0
C.Middle	3.4
C.High	3.8
D.Low	2.8

STEP 4: The sub-MFFP tree for two different quantitative databases are respectively built. The results of two trees are respectively shown in Figures 2 and 3.

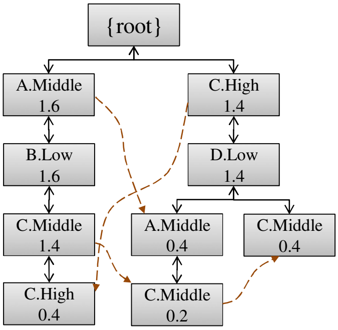


Fig. 2. The sub-MFFP tree of DB_1

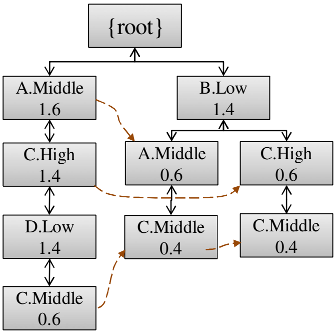


Fig. 3. The sub-MFFP tree of DB_2

STEP 5: The leaf nodes of MFFP-tree in DB_1 are then traced one by one. In this example, three branches can be desired from these leaf nodes. Figure 4 shows the result of the three branches.

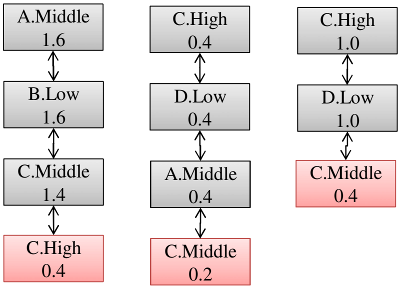
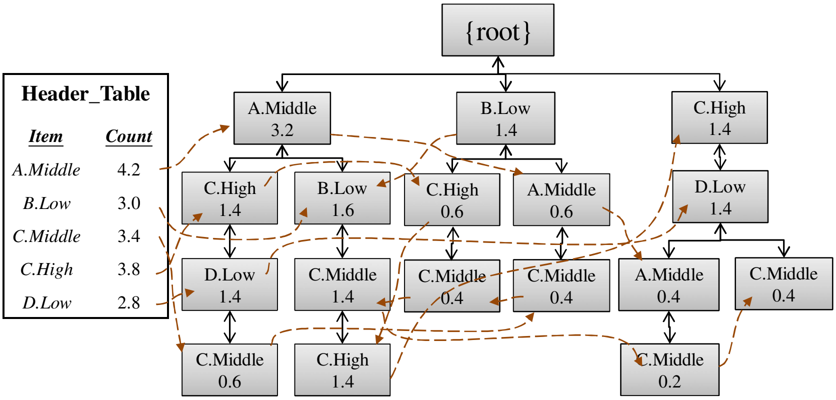


Fig. 4. Three branches of the currently sub-MFFP tree

STEP 6: Insert the three branches of the sub-MFFP tree of DB_1 into the iMFFP tree. Note that the sub-MFFP tree of DB_2 is considered as the initial iMFFP-tree in this example.

STEPs 7 & 8: After the step 6 is executed, two sub-MFFP trees are then merged together. Since no sub-MFFP trees should be merged, we create the Header_Table and insert node-links from the entry of a fuzzy region in the Header_Table to the first branch of that fuzzy region. The final iMFFP-tree has thus been constructed. The final results are shown in Figure 5.


Header_Table

Item	*Count*
A.Middle	4.2
B.Low	3.0
C.Middle	3.4
C.High	3.8
D.Low	2.8

Fig. 5. The final iMFFP-tree

5 Experimental Results

The experiments were performed on a real dataset called CONNECT [4]. It was divided into 2 and 5 sub-databases for constructing 2 and 5 sub-MFFP trees. The execution time of the proposed iMFFP-tree algorithm and the PFPTC algorithm [14] was compared in different minimum support thresholds. The results are shown in Figure 6.

It is obvious to see that the proposed algorithm had a better performance than the PFPTC algorithm both in the two or five sub-databases for integration.

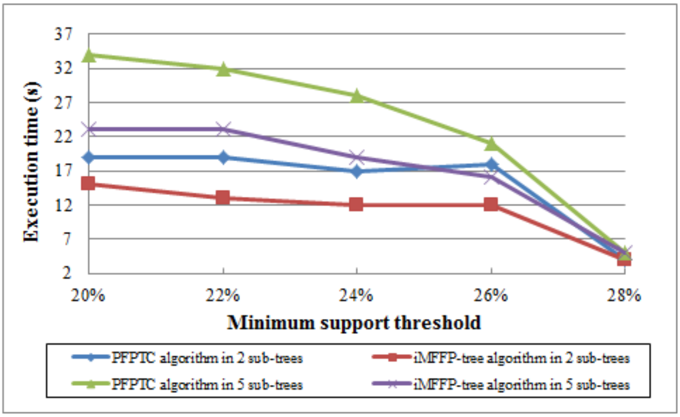


Fig. 6. The comparisons of execution times

6 Conclusions

In real-world applications, the information from several branches can be integrated into efficient rules for a parent industry to make correct decision. Thus, we propose the integrating MFFP-tree (iMFFP-tree) algorithm for merging several sub-MFFP trees into one. The branches in the sub-MFFP tree are efficiently extracted to integrate the other sub-MFFP trees in a sequence, thus forming an integrated MFFP tree for decision making. Experimental results also show that the proposed iMFFP-tree algorithm has a better performance than the PFPTC algorithm.

References

1. Agrawal, R., Srikant, R.: Fast algorithms for mining association rules in large databases. In: The International Conference on Very Large Data Bases, pp. 487–499 (1994)
2. Agrawal, R., Srikant, R.: Mining sequential patterns. In: The International Conference on Data Engineering, pp. 3–14 (1995)
3. Agrawal, R., Imielinski, T., Swami, A.: Mining association rules between sets of items in large databases. In: The International Conference on Management of Data, pp. 207–216 (1993)
4. Bayardo, R.: UCI repository of machine learning databases, `http://fimi.ua.ac.be/data/connect.dat`
5. Berzal, F., Cubero, J.C., Marín, N., Serrano, J.M.: Tbar: An efficient method for association rule mining in relational databases. Data and Knowledge Engineering 37, 47–64 (2001)
6. Chen, M.S., Han, J., Yu, P.S.: Data mining: An overview from a database perspective. IEEE Transactions on Knowledge and Data Engineering 8, 866–883 (1996)
7. Han, J., Pei, J., Yin, Y., Mao, R.: Mining frequent patterns without candidate generation: A frequent-pattern tree approach. Data Mining and Knowledge Discovery 8, 53–87 (2004)

8. Hong, T.P., Kuo, C.S., Wang, S.L.: A fuzzy aprioritid mining algorithm with reduced computational time. Applied Soft Computing 5, 1–10 (2004)
9. Hong, T.P., Lin, C.W., Lin, T.C.: Mining complete fuzzy frequent itemsets by tree structures. In: IEEE International Conference on Systems, Man, and Cybernetics, pp. 563–567 (2010)
10. Hu, K., Lu, Y., Zhou, L., Shi, C.: Integrating classification and association rule mining: A concept lattice framework. In: The International Workshop on New Directions in Rough Sets, Data Mining, and Granular-Soft Computing, pp. 443–447 (1999)
11. Kuok, C.M., Fu, A., Wong, M.H.: Mining fuzzy association rules in databases. SIGMOD Record 27, 41–46 (1998)
12. Lent, B., Swami, A., Widom, J.: Clustering association rules. In: The International Conference on Data Engineering, pp. 220–231 (1997)
13. Liu, F., Lu, Z., Lu, S.: Mining association rules using clustering. Intelligent Data Analysis 5, 309–326 (2001)
14. Lan, Y.J., Qiu, Y.: Parallel frequent itemsets mining algorithm without intermediate result. In: The International Conference on Machine Learning and Cybernetics, pp. 2102–2107 (2005)
15. Papadimitriou, S., Mavroudi, S.: The fuzzy frequent pattern tree. In: The WSEAS International Conference on Computers, pp. 1–7 (2005)
16. Park, J.S., Chen, M.S., Yu, P.S.: Using a hash-based method with transaction trimming for mining association rules. IEEE Transactions on Knowledge and Data Engineering 9, 813–825 (1997)
17. Srikant, R., Agrawal, R.: Mining sequential patterns: Generalizations and performance improvements. In: The International Conference on Extending Database Technology: Advances in Database Technology, pp. 3–17 (1996)
18. Sucahyo, Y., Gopalan, R.: Building a More Accurate Classifier Based on Strong Frequent Patterns. In: Webb, G.I., Yu, X. (eds.) AI 2004. LNCS (LNAI), vol. 3339, pp. 1036–1042. Springer, Heidelberg (2004)
19. Zadeh, L.A.: Fuzzy sets. Information and Control, 338–353 (1965)

A Quadratic Algorithm for Testing of Omega-Codes

Nguyen Dinh Han[1], Phan Trung Huy[2], and Dang Quyet Thang[3]

[1] Hung Yen University of Technology and Education, Vietnam
hannguyen@utehy.edu.vn
[2] Hanoi University of Science and Technology, Vietnam
phanhuy@hn.vnn.vn
[3] Nam Dinh University of Technology and Education, Vietnam
thangdgqt@gmail.com

Abstract. We consider a special class of codes, namely ω-codes related to infinite word which had been studied by many authors. Until now, the best algorithm to test whether a regular language X is an ω-code has time complexity $\mathcal{O}(n^3)$, where n is the size of the transition monoid of the minimal automaton recognizing X. In this paper, with any monoid M saturating X (the transition monoid above is only a special case), we propose a new test and establish a quadratic testing algorithm with time complexity $\mathcal{O}(n^2)$ to verify if X is an ω-code, where n is Card(M).

Keywords: infinitary, quadratic algorithm, infinite word, finite monoid, omega-code.

1 Preliminaries

The notion of infinitary codes has been considered in [10,9]. In the class of codes of finite words, a subclass of ω-codes has been studied in many works [7,6,4,1] which showed its role in development of languages of finite and infinite words. Until now, as shown in a deep survey on tests for ω-codes conducted by Augros and Litovsky [1], the best testing algorithm for ω-codes has time complexity $\mathcal{O}(n^3)$, where n is the size of the transition monoid of the minimal automaton recognizing the input language. In Section 2 of our paper, a new technique based on finite graphs to establish a test for ω-codes is presented. As a consequence, we obtain an effective testing algorithm for ω-codes with time complexity $\mathcal{O}(n^2)$ which is the main result of the paper and it is presented in Section 3.

In the following, we recall some notions (for more details, we refer to [3,5]). Let A be a finite alphabet. As usual, A^* is the free monoid generated by A whose elements are called *finite words*, A^ω is the set of all *infinite words*. An infinitary language can include both finite and infinite words, the largest $A^\infty = A^* \cup A^\omega$. The empty word is denoted by ε and $A^+ = A^* - \{\varepsilon\}$. The length $|w|$ of the word $w = a_1 a_2 \cdots a_n$ with $a_i \in A$ is n. By convention $|\varepsilon| = 0$. A subset of A^* is called a *language* while a subset of A^ω is an *ω-language*. For any language $X \subseteq A^*$, we denote by X^* the submonoid of A^* generated by X, $X^+ = X^* - \{\varepsilon\}$ and X^ω the ω-language

J.-S. Pan, S.-M. Chen, N.T. Nguyen (Eds.): ACIIDS 2012, Part I, LNAI 7196, pp. 338–347, 2012.

$$X^\omega = \{w \in A^\omega \mid w = u_1u_2 \cdots \text{with}\, u_i \in X\}.$$

A *factorization* of a word w in X^+ (respectively in X^ω) is a finite sequence (resp. an infinite sequence) $\{u_1, u_2, \ldots, u_n\}$ (resp. $\{u_1, u_2, \ldots\}$) of words of X such that $w = u_1u_2 \cdots u_n$, $n \geq 1$ (resp. $w = u_1u_2 \cdots$). A language $X \subseteq A^+$ is a code (resp. an ω-code) if every word w in A^* (resp. in A^ω) has at most one factorization on X.

Let M be a monoid. For $S, T \subseteq M$, we define *left quotients and right quotients* of S by T as follows

$$T^{-1}S = \{u \in M \mid \exists t \in T, t.u \in S\},$$

$$ST^{-1} = \{u \in M \mid \exists t \in t, u.t \in S\}.$$

The notations $u^{-1}S, Su^{-1}$ will be used when $T = \{u\}$ is singleton. For any $u, v \in M$, we write uv instead of $u.v$ whenever $M = A^*$.

Given $X \subseteq A^*$, we say that X *is saturated by a monoid morphism* $\varphi : A^* \to M$ if there exists $B \subseteq M$ such that $X = \varphi^{-1}(B)$ and in that case, we also say that M saturates X and X is given by this tuple (φ, M, B). In case X is regular, M can be chosen by the transition monoid of the minimal automaton recognizing X, or by the syntactic monoid M_X. We say that $k = \text{Card}(M_X)$ *is the index of the syntactic congruence of* X, briefly k *is the index of* X or X *of index* k.

The following result is well-known for regular languages.

Lemma 1. *Let $X \subseteq A^*$ be a language. The following conditions are equivalent:*
(*i*) *X is regular.*
(*ii*) *X is recognized by a finite automaton* (*or X is recognizable*)*.*
(*iii*) *X is saturated by a morphism from A^* onto a finite monoid M.*

Since X is a code if and only if X^* is a free monoid, the code properties of X can be verified on the submonoid generated by X. The following basic properties that can be checked by definition, provide us tools to establish an effective algorithm to test whether a regular language of finite words is an ω-code or not.

Lemma 2. *Let $h : A^* \to M$ be a surjective monoid morphism saturating X, Y. Suppose $X = h^{-1}(K)$, $Y = h^{-1}(L)$, $h(X^+) = T$ for some $K, L, T \subseteq M$, then $X \cup Y = h^{-1}(K \cup L)$, $X \cap Y = h^{-1}(K \cap L)$, $X - Y = h^{-1}(K - L)$, $X^{-1}Y = h^{-1}(K^{-1}L)$, $XY^{-1} = h^{-1}(KL^{-1})$, and $(X^+)^{-1}Y = h^{-1}(T^{-1}L)$.*

Proof. By assumption $X = h^{-1}(K), Y = h^{-1}(L)$, using basic property of mapping we directly obtain $X \cup Y = h^{-1}(K \cup L)$, $X \cap Y = h^{-1}(K \cap L)$, $X - Y = h^{-1}(K - L)$, $X^{-1}Y = h^{-1}(K^{-1}L)$, $XY^{-1} = h^{-1}(KL^{-1})$.

To prove $(X^+)^{-1}Y = h^{-1}(T^{-1}L)$, since $h(X^+) = T$ and $X = h^{-1}(K)$, we have $K^+ = T$.

Now, for all $w \in (X^+)^{-1}Y$, there exist $x_1, x_2, \ldots, x_n \in X$, $y \in Y$ such that $x_1x_2 \cdots x_nw = y$. This implies $h(x_1).h(x_2) \cdots h(x_n).h(w) = h(y) \in L$. Since $h(x_i) \in K$ then $h(w) \in (K^+)^{-1}L$. Thus $w \in h^{-1}(T^{-1}L)$.

Conversely, let $w \in h^{-1}(T^{-1}L)$, we have

$$h(w) \in T^{-1}L = (K^+)^{-1}L \Leftrightarrow \exists \alpha \in K^+ : \alpha.h(w) \in L.$$

Now $\alpha \in K^+$ implies $\alpha = \alpha_1.\alpha_2 \cdots \alpha_p, \alpha_i \in K, i = 1, \ldots, p$. Since h is surjective and $X = h^{-1}(K)$, then there exist $x_i \in X, i = 1, \ldots, p$ such that $\alpha_i = h(x_i)$. We have $\alpha = h(x_1x_2 \cdots x_p)$, then $\alpha.h(w) = h(x_1x_2 \cdots x_pw)$. This implies $x_1x_2 \cdots x_pw \in h^{-1}(L) = Y$ and we have $w \in (X^+)^{-1}Y$. □

By Lemma 2, the following corollary is obvious

Corollary 1. *Let $X, Y \subseteq A^*$ be regular languages, and let P be a finite monoid, and $h : A^* \to P$ be a surjective monoid morphism. If X and Y are saturated by h, then h saturates all $L \in \mathcal{A}(X, Y)$, where $\mathcal{A}(X, Y)$ is the class of languages obtained from X, Y by taking a finite number of boolean operations and formation of left and right quotients.*

Proposition 1. *Let $X, Y \subseteq A^*$ be regular languages, and $f : A^* \to M$ and $g : A^* \to N$ be monoid morphisms, f saturates X and g saturates Y. Let $h : A^* \to P \subseteq M \times N$ be the surjective monoid morphism given by: $\forall a \in A^*$, $h(a) = (f(a), g(a))$ and P is the submonoid of $M \times N$ generated by the set of all elements $h(a), a \in A$. Then h saturates both X and Y.*

Example 1. (i) Let $\varphi : A^* \to M$ be a monoid morphism saturating $X, X \subseteq A^*, A \neq \emptyset$, and let $Y = \{\varepsilon\}$. Consider $Z_2 = \{0, 1\}$ as the monoid of integers modulo 2 with the unit 1. Obviously, $Y = \{\varepsilon\}$ is saturated by the monoid morphism $g : A^* \to Z_2$, which is defined by: $g(\varepsilon) = 1$ and for all $a \in A$, $g(a) = 0$. It is easy checked that both X and $Y = \{\varepsilon\}$ are saturated by the surjective monoid morphism $h : A^* \to P \subseteq M \times Z_2$ defined by: $\forall a \in A^*$, $h(a) = (\varphi(a), 0)$, $1_P = h(\varepsilon) = (1_M, 1)$.

(ii) In the case $Y = \{\varepsilon\}$, a more simple method for constructing a surjective morphism saturating both X and Y can be presented as follows

Extend M to new monoid $M^1 = M \cup \{\mathbf{1}\}$ where $\mathbf{1} \notin M$ as the new unit of M^1 by extra conditions $(\mathbf{1}.x = x.\mathbf{1} = x, \forall x \in M^1)$.

Define $h' : A^* \to M^1$ by $h'(a) = \varphi(a), \forall a \in A$ but $h'(\varepsilon) = \mathbf{1}$.

From $h' : A^* \to M^1$ we define the surjective morphism $h_1 : A^* \to P = h'(A)^* \subseteq M^1$ which is induced from h': $h_1(a) = h'(a), \forall a \in A, h_1(\varepsilon) = h'(\varepsilon) = \mathbf{1}$.

2 Testing Rational ω-Codes

2.1 A Language Algorithm

In this section, we recall a test for ω-codes proposed in [1].

Given a language $X \subseteq A^+$, we consider a sequence of sets V_i defined recursively as follows

$$V_1 = (X^+)^{-1}X - \{\varepsilon\}, \quad V_{i+1} = (V_iX^*)^{-1}X, \quad i \geq 1. \qquad (1)$$

Then, we have the following theorem that gives a criterion for ω-codes.

Theorem 1 ([1]). *Let $X \subseteq A^+$ be a regular language and let $V_i\,(i \geq 1)$ be defined in (1). Then, X is an ω-code if and only if there exists $i \geq 1$ such that $V_i = \emptyset$.*

Remark 1. The correctness of the test is deduced from Theorem 1, but we remark that the proof of Theorem 1 in [1] should be improved with the use of lemmata and propositions.

We deduce the following direct result

Corollary 2. *Let $X \subseteq A^+$ and let $V_i\,(i \geq 1)$ be defined in (2). If $\varepsilon \in V_i$ for some $i \geq 1$ then X is not an ω-code.*

Example 2. Let $A = \{a, b\}$.

For $X = \{aa, baa, ba\}$, we have $V_1 = V_2 = \cdots = \{a\}$. This implies $V_i \neq \emptyset$ for all $i \geq 1$, thus X is not an ω-code.

For $Y = \{a, b\}$, we have $V_1 = \emptyset$, thus Y is an ω-code.

2.2 An Algorithm Basing on Monoids

Given a regular language $X \subseteq A^*$, then by Proposition 1, X is saturated by a surjective morphism h from A^* onto a finite monoid P. As shown in [1], by using monoid approach, one can build an $\mathcal{O}(n^3)$ algorithm to test whether X is an ω-code, where n is the size of the transition monoid of the minimal automaton recognizing X. In this section, with n is the size of any monoid P saturating X, we establish a quadratic algorithm to test whether X is an ω-code or not. In Theorem 2 below, we show that verifying whether X is an ω-code reduces to verifying whether there exists *a special cycle* in P.

Definition 1. *Let $X \subseteq A^+$ be a regular language and let $h : A^* \to P$ be a surjective monoid morphism saturating both $Y = \{\varepsilon\}$ and X, where P is finite, $X = h^{-1}(K)$, $Y = h^{-1}(L)$, $L = \{1_P\}, S = K^*, T = K^+$ with $K, L \subseteq P$. A rho-cycle in P is a sequence $e_1, e_2, \ldots, e_j$ such that $e_1, e_2, \ldots, e_{j-1}$ are all distinct, but $e_j = e_i, j > i$ with*

$$e_1 \in T^{-1}K - L, \quad e_{i+1} \in (e_i.S)^{-1}K, \quad i \geq 1. \tag{2}$$

Then, we have the following results

Lemma 3. *Let $X \subseteq A^+$ be a regular language and let $h : A^* \to P$ be a surjective monoid morphism saturating both $Y = \{\varepsilon\}$ and X, where P is finite, $X = h^{-1}(K)$, $Y = h^{-1}(L)$, $L = \{1_P\}, S = K^*, T = K^+$ with $K, L \subseteq P$. Let $V_i\,(i \geq 1)$ be defined in (1) and let $e_1, e_2, \ldots, e_j = e_i, j > i$ be a rho-cycle in P. Then, we have $h^{-1}(e_n) \subseteq V_n$ for all $n \geq 1$.*

Proof. The proof is proceeded by induction on n.

For $n = 1$. First, let $\alpha \in h^{-1}(e_1)$, then $h(\alpha) = e_1 \in T^{-1}K - L$. We have

$$h(\alpha) \in T^{-1}K - L = (K^{+})^{-1}K - \{1_P\} \Leftrightarrow (\exists p \in K^{+}, k \in K : h(\alpha) \in p^{-1}k, h(\alpha) \neq 1_P \Leftrightarrow k = p.h(\alpha), h(\alpha) \neq 1_P).$$

Now $p \in K^{+}$ implies that that $p = k_1.k_2.\cdots.k_\lambda, k_i \in K, i = 1, \ldots, \lambda$. Since h is surjective and $X = h^{-1}(K)$, then there exist $x \in X$ such that $k = h(x)$ and $y_i \in X$ such that $k_i = h(y_i)$, $i = 1, \ldots, \lambda$. Thus, we have

$$h(x) = h(y_1).h(y_2).\cdots.h(y_\lambda).h(\alpha) = h(y_1y_2\cdots y_\lambda\alpha) \in K, h(\alpha) \neq 1_P.$$

This implies $u = y_1y_2\cdots y_\lambda\alpha \in h^{-1}(K) = X$. Thus $\alpha = (y_1y_2\cdots y_\lambda)^{-1}u \in (X^{+})^{-1}X$. Since $h(\alpha) \neq 1_P$, then $\alpha \neq \varepsilon$. Hence $\alpha \in (X^{+})^{-1}X - \{\varepsilon\} = V_1$.

Hence the assertion holds for $n = 1$.

Suppose now it is true for some $n > 1$, we prove it remains true for $n + 1$. First, let $\alpha \in h^{-1}(e_{n+1})$ and let $\gamma \in h^{-1}(e_n)$, then $h(\alpha) \in (e_n.S)^{-1}K$. We have

$$h(\alpha) = e_{n+1} \in (e_n.S)^{-1}K \Leftrightarrow (\exists p \in K^{*}, k \in K : h(\alpha) \in (e_n.p)^{-1}k \Leftrightarrow k = e_n.p.h(\alpha)).$$

This implies $h(x) = h(\gamma y_1y_2\cdots y_\lambda\alpha) \in K$ with $x \in X$ and $y_1, y_2, \ldots, y_\lambda \in X$, $\lambda \geq 0$. We have $u = \gamma y_1y_2\cdots y_\lambda\alpha \in X$. By assumption, $\gamma \in h^{-1}(e_n) \subseteq V_n$. Hence $\alpha = (\gamma y_1y_2\cdots y_\lambda)^{-1}u \in (V_nX^{*})^{-1}X = V_{n+1}$.

Therefore, the assertion holds for all $n \geq 1$. □

Then, we have the following theorem that gives a criterion for ω-codes.

Theorem 2. *Let $X \subseteq A^{+}$ be a regular language and let $h : A^{*} \to P$ be a surjective monoid morphism saturating both $Y = \{\varepsilon\}$ and X, where P is finite, $X = h^{-1}(K)$, $Y = h^{-1}(L)$, $L = \{1_P\}, S = K^{*}, T = K^{+}$ with $K, L \subseteq P$. Let V_i $(i \geq 1)$ be defined in (1). Then, X is not an ω-code if and only if there exists a rho-cycle $e_1, e_2, \ldots, e_j = e_i, j > i$ in P.*

Proof. ($\Rightarrow$) Suppose that X is not an ω-code, then by Theorem 1, $V_i \neq \emptyset$ for all i. Let $N = \mathrm{Card}(P) + 1$, then the exists a sequence $v_1, v_2, \ldots, v_{N-1}, v_N$ with $v_1 \in V_1$, $v_{i+1} \in (v_iX^{*})^{-1}X$, $N > i \geq 1$. Indeed, we construct this sequence as follows. First, since $V_N \neq \emptyset$, we can always pick out $v_N \in V_N = (V_{N-1}X^{*})^{-1}X$. Next, we have

$$v_N \in (V_{N-1}X^{*})^{-1}X \Leftrightarrow (\exists x \in X^{*}, v_{N-1} \in V_{N-1}, x_1 \in X : v_N = (v_{N-1}x)^{-1}x_1 \in (v_{N-1}X^{*})^{-1}X).$$

Thus we can always pick out $v_{N-1} \in V_{N-1}$ such that $v_N \in (v_{N-1}X^{*})^{-1}X$. We apply this argument until we obtain v_1.

Now, let $e_i \in h(v_i)$ we show that there exists a rho-cycle $e_1, e_2, \ldots, e_j = e_i, j > i$ in P. First, let $e_1 \in h(v_1)$. We have

$$v_1 \in V_1 \Leftrightarrow (\exists x_1x_2\cdots x_\lambda \in X^{+}, x \in X, \lambda \geq 1, x \in X : x_1x_2\cdots x_\lambda v_1 = x, v_1 \neq \varepsilon).$$

Since h is surjective, and $X = h^{-1}(K)$ then there exist $k_1, k_2, \ldots, k_\lambda \in K$ such that $k_i = h(x_i), \lambda \geq i \geq 1$ and $k \in K$ such that $k = h(x)$. Thus, we have $k_1.k_2.\cdots.k_\lambda.e_1 = k \in K, e_1 \neq 1_P$. Hence $e_1 \in (k_1.k_2.\cdots.k_\lambda)^{-1}k \subseteq (K^{+})^{-1}$ $K - L$.

Next, let $e_2 \in h(v_2)$, there exist $x_1, x_2, \ldots, x_\lambda, x \in X$ and $k_1, k_2, \ldots, k_\lambda, k \in K$, $\lambda \geq 0$ such that $k_i = h(x_i), k = h(x)$ and $e_1.k_1.k_2.\cdots.k_\lambda.e_2 = k \in K$. Therefore $e_2 \in (e_1.k_1.k_2.\cdots.k_\lambda)^{-1}k \subseteq (e_1.S)^{-1}K$.

We apply this argument until we obtain e_N.

If $e_1, e_2, \ldots, e_N$ are all distinct then the number of different e_i's exceeds Card(P). Thus, there exist $i < j \leq N$ such that $e_j = e_i$. Therefore P has a rho-cycle $e_1, e_2, \ldots, e_j = e_i, j > i$.

($\Leftarrow$) Suppose that P has a rho-cycle $e_1, e_2, \ldots, e_j = e_i, j > i$. Now we put an arrow $e_i \to e_{i+1}$ if $e_{i+1} \in (e_i.S)^{-1}K$. Then, by assumption, we have a path

$$e_1 \to e_2 \to \cdots \to e_i \to e_{i+1} \to \cdots \to e_j$$

We have $e_{i+1} \in (e_i.S)^{-1}K = (e_j.S)^{-1}K$ since $e_j = e_i$. Thus, we have a path

$$e_1 \to e_2 \to \cdots \to e_i \to e_{i+1} \to \cdots \to e_j \to e_{i+1} \tag{3}$$

which implies an infinite path

$$e_1 \to e_2 \to \cdots \to e_i \to e_{i+1} \to \cdots \to e_j \to e_{i+1} \to \cdots$$

Thus, there exists an infinite sequence $e_1, e_2, \ldots, e_j = e_i, e_{j+1} = e_{i+1}, \ldots$ Then, by Lemma 3, $\emptyset \neq h^{-1}(e_i) \subseteq V_i$, $\forall i \geq 1$. We deduce that $V_i \neq \emptyset$ for all $i \geq 1$. This implies that X is not an ω-code.

The proof is completed. □

2.3 A Graph Algorithm

In this section, our main consideration is a graph algorithm for testing of ω-codes. From the monoid P and $K, S \subseteq P$ defined in Section 2.2, we construct a coloured directed graph $G = (V, E)$ as follows. Firstly, $E = \emptyset$. Let $V = P$ and E is updated by

If $m \in e.S$, then we add a *red arrow* $e \xrightarrow{red} m$ into E.

If $a.b \in K$ or equivalently $b \in a^{-1}K$, then we add a *blue arrow* $a \xrightarrow{blue} b$ into E.

We have the fact

$$e \xrightarrow{red} m \xrightarrow{blue} e' \Leftrightarrow (m \in e.S \wedge m.e' \in K) \Leftrightarrow e' \in (e.S)^{-1}K. \tag{4}$$

Now for each path mentioned in (3)

$$e_1 \to e_2 \to \cdots \to e_i \to e_{i+1} \to \cdots \to e_j \to e_{i+1}$$

can be extended to a path in G

$$e_1 \xrightarrow{red} m_1 \xrightarrow{blue} e_2 \xrightarrow{red} \cdots \xrightarrow{blue} e_{i+1} \xrightarrow{red} \cdots \xrightarrow{blue} e_j \xrightarrow{red} m_j \xrightarrow{blue} e_{i+1}.$$

Then, as a consequence of Theorem 2, the following theorem is obvious

Theorem 3. *Let $X \subseteq A^+$ be a regular language and $G = (V, E)$ be defined as above. Then, X is not an ω-code if and only if there exists a rho-cycle $e_1, e_2, \ldots, e_j = e_i, j > i$ in G.*

3 A Quadratic Algorithm for Testing of ω-Codes

At first, by considering a path

$$e_1 \to e_2 \to \cdots \to e_i \to e_{i+1} \to \cdots \to e_j$$

and using the cycle detection algorithm on G [2,8], it seems that we can obtain a simple algorithm to test for ω-codes. But unfortunately, adding arrows $e_i \to e_{i+1}$ by condition ($e_i \to e_{i+1}$ if $e_{i+1} \in (e_i.S)^{-1}K$) can lead to an algorithm running in $\mathcal{O}(n^3)$ time, which is not better than the one in [1]. To avoid this, by a trick expanding G to a new graph G' so that we can apply the algorithm [2,8] to obtain a quadratic algorithm for testing of ω-codes.

We can assume the input of this algorithm as a tuple (φ, M, B), with $\varphi : A^* \to M$ is a monoid morphism saturating X, M is a finite monoid, $B \subseteq M$, $X = \varphi^{-1}(B)$, and the claim for the algorithm is the answer whether or not X is an ω-code and suppose that $\mathrm{Card}(A)$ is a constant.

Lemma 4. *Let M be monoid,* $\mathrm{Card}(M) = n$*, then there exists an algorithm to find K^*, K^+ for any $K \subseteq M$, with time complexity $\mathcal{O}(n^2)$.*

Proof. Indeed, we have $K^+ = K^*.K$,

$$K^* = \begin{cases} K^+ & \text{if } 1_M \in K^+ \\ K^+ \cup 1_M & \text{otherwise.} \end{cases}$$

Now, for presentation of $K, T = K^+, S = K^*$ as subsets of M, we use some arrays of n elements. For every $e \in M$, we present

$$flagK(e) = 1 \Leftrightarrow e \in K,$$
$$flagS(e) = 1 \Leftrightarrow e \in S,$$
$$flagT(e) = 1 \Leftrightarrow e \in T.$$

Step 0: Initiation, for all $e \in M$ we set $flagK(e) = flagS(e) = flagT(e) = 0$. The time and space complexity for this step is $\mathcal{O}(n)$.
Step 1 (Update the array $flagK$): For every $e \in K$, set $flagK(e) = 1$. The time complexity for this step is $\mathcal{O}(n)$.
Step 2: We need to define a table T for presentation of the binary operation on M: for every $e, f \in M$, $e.f = m$ if and only if $T(e, f) = m$. This step can be completed in time and space complexity $\mathcal{O}(n^2)$.
Step 3: Constructing a directed graph $G = (V, E)$. At first, $V = M, E = \emptyset$ and E is updated by

```
for $x \in M$ do
   for $f \in M$ do
       if $e \neq f$ and $T(f, x) = e$ then
          add $f \longrightarrow e$ into $E$;
```

Hence, Step 3 can be done in time and space complexity $\mathcal{O}(n^2)$.
Step 4 (Update the array $flagT$): It is easy to see that for every $m \in M$,

$$m \in K^+ \Leftrightarrow m \in 1_M.K^+ \Leftrightarrow (\exists e_1, e_2, \ldots, e_l \in M : e_1 \in 1_M.K, e_2 \in e_1.K, e_3 \in e_2.K, \cdots, m = e_l \in e_{l-1}.K),$$

or equivalently, there exists a path starting from 1_M and ending at m in G. Using the Dijkstra algorithm, we can construct a subprocedure $visit(1_M)$ with time complexity $\mathcal{O}(n^2)$ to update the array $flagT$ (i.e. $flagT(m) = 1 \Leftrightarrow m \in T$).
Step 5 (Update the array $flagS$): Two cases happen

If $1_M \in K^+$, or equivalently, $flagT(1_M) = 1$ then set $flagS(e) = flagT(e)$ for any $e \in M$.

Otherwise, $1_M \notin K^+$, then set $flagS(e) = flagT(e)$ for any $1_M \neq e \in M$ and $flagS(1_M) = 1$. This step has time complexity $\mathcal{O}(n)$.

Hence, totally, the time and space complexity for the whole 5 steps is $\mathcal{O}(n^2)$. This completes the proof. □

Remark 2. As a direct consequence of Lemma 4, from a given monoid morphism $h : A^* \to M$ which saturates a regular language $X \subseteq A^*$, $\text{Card}(M) = n$, we can define $P = K^*$ from $K = h(A)$ with time complexity $\mathcal{O}(n^2)$. Use this P, we can define a surjective morphism $h' : A^* \to P$ which also saturates X.

Theorem 4. *Let $X \subseteq A^+$ and let $h : A^* \to P$ be a surjective monoid morphism saturating X. There exists an algorithm with time complexity $\mathcal{O}(n^2)$ to determine whether there exists any rho-cycle in P.*

Proof. Indeed, we can assume X is given by a tuple (h, K, P) with $h^{-1}(K) = X$.
Step 1 (constructing G).

From P, K, we can construct $S = K^*, T = K^+$ with the time complexity $\mathcal{O}(n^2)$ by using Lemma 4.

Use S, T, K, P, we define a graph $G = (V, E)$, with $V = P \times I$ where $I = \{1, 2\}$, and E is constructed by the following procedure

```
for $e \in P$ do
   for $s \in P$ do
       if $s \in S$ (or equivalently $flagS(s) == 1$) then
          add the arrow $((e,1) \xrightarrow{red} (e.s,2))$ to $E$;
for $m \in P$ do
   for $e \in P$ do
       if $m.e \in K$ (or equivalently $flagK(m.e) == 1$) then
          add the arrow $((m,2) \xrightarrow{blue} (e,1))$ to $E$
       if $m.e \in K$ and $m \in T$ and $e \neq 1_P$ then
          $start(e,1) = 1$;
```

Note that we present $start(e, 1) = 1$ for the condition $e \in T^{-1}K - L$ where the *start* is an array of n elements in P.

Hence, constructing G requires time $\mathcal{O}(n^2)$.

Step 2 (detecting a rho-cycle in P is equivalent to detecting a rho-cycle in G)

```
boolean containsCycle(Graph g)
        for each vertex v in g do
            v.mark = WHITE;
        for each vertex v in g do
            if v.mark == WHITE and start(v, 1) == 1 then
              if visit(g, v) then
                return TRUE;
        return FALSE;

boolean visit(Graph g, Vertex v)
        v.mark = GREY;
        for each edge (v, u) in g do
            if u.mark == GREY and u.index == 1 then
              return TRUE;
            else if u.mark == WHITE then
                if visit(g, u) then
                  return TRUE;
        v.mark = BLACK;
        return FALSE;
```

Note that we use the notation $u.index = 1$ to denote $(u, 1) \in V$.

Let us remark that the algorithm presented in Step 2 is called coloured depth first search (DFS) algorithm whose time complexity is $\mathcal{O}(\mathrm{Card}(V) + \mathrm{Card}(E))$ [2,8]. We have $\mathrm{Card}(V) = 2.\mathrm{Card}(P) \leq 2.(\mathrm{Card}(M) + 1) = 2n + 2$, $\mathrm{Card}(E) \leq (2n+2)^2$. Hence, Step 2 can be completed in $\mathcal{O}(n^2)$. □

Algorithm 1. (*The test for ω-codes in regular case*)

```
Input:   A regular language $X \subseteq A^+$ is given by a tuple $(\varphi, M, B)$,
         with $\varphi : A^* \to M$, $M$ is a finite monoid, $B \subseteq M$, $X = \varphi^{-1}(B)$.
Output: "YES" if $X$ is an $\omega$-code, "NO" otherwise.
Step 1:  From the tuple $(\varphi, M, B)$, construct a surjective monoid morphism
         $h : A^* \to P \subseteq M^1$ saturating both $X$ and $Y = \{\varepsilon\}$
         (by the method in Example 1 (ii))
Step 2:  Calculate $K = P \cap B$ which satisfies $h^{-1}(K) = X$.
Step 3:  Calculate $T = K^+, S = K^*$.
Step 4: From $P, K, S, T$ construct a directed graph $G$ with the set of vertices is
         $P \times I$ by the method mentioned in the proof of Theorem 4 (Step 1).
Step 5:  Determine whether there exists a rho-cycle in $G$
         If there exist some rho-cycles in $G$ then
             Return "NO";
         Else
             Return "YES";
```

Details and Complexity of the Algorithm

1) In Step 1, by using Lemma 4, as in Example 1 (*ii*), the time complexity is $\mathcal{O}(n^2)$ to define two arrays $flagP$, $flagB$.

2) In Step 2, calculating K requires a time complexity $\mathcal{O}(n)$ by conditions: $flagK(x) = 1$ if and only if $flagP(x) = 1$ and $flagB(x) = 1$ for any $x \in M^1$.

3) In Step 3, calculating $T = K^+$ and $S = K^*$ as subsets of P require a time complexity $\mathcal{O}(n^2)$, by Lemma 4.

4) In Step 4 and Step 5, applying Theorem 4, the time complexity for these steps is $\mathcal{O}(n^2)$.

As a consequence of Theorems 3 and 4, the main result of this section can be formulated as follows

Theorem 5. *Given as input a tuple* (φ, M, B), *with* $\varphi : A^* \to M$ *is a monoid morphism saturating* X, M *is a finite monoid,* $B \subseteq M$, $X = \varphi^{-1}(B)$, *Algorithm 1 allows us to decide whether or not* X *is an* ω*-code in* $\mathcal{O}(n^2)$ *time, where* $n = \mathrm{Card}(M)$.

Acknowledgment. The authors acknowledge their great indebtedness to the referees for their valuable comments and suggestions.

References

1. Augros, X., Litovsky, I.: Algorithm to test rational ω-codes. In: Proceedings of the Conference of The Mathematical Foundation of Informatics, pp. 23–37. World Scientific (October 1999)
2. Berman, K.A., Paul, J.L.: Algorithms - Sequential, parallel, and distributed. Thomson Learning, Inc., USA (2005)
3. Berstel, J., Perrin, D., Reutenauer, C.: Theory of Codes. Academic Press Inc., New York (1985)
4. Devolder, J., Latteux, M., Litovsky, I., Staiger, L.: Codes and infinite words. Acta Cybernetica 11(4), 241–256 (1994)
5. Lallement, G.: Semigroups and Combinational Applications. John Wiley and Sons, Inc. (1979)
6. Lam, N.H., Van, D.L.: On strict codes. Acta Cybernetica 10(1-2), 25–34 (1991)
7. Mateescu, A., Mateescu, G.D., Rozenberg, G., Salomaa, A.: Shuffle-like operations on ω-words. In: Păun, G., Salomaa, A. (eds.) New Trends in Formal Languages. LNCS, vol. 1218, pp. 395–411. Springer, Heidelberg (1997)
8. Sedgewick, R.: Algorithms in C++, Part 5: Graph algorithms. Addition-Wesley, Pearson Education, Inc., USA (2002)
9. Staiger, L.: On infinitary finite length codes. Informatique Théorique et Applications 20(4), 483–494 (1986)
10. Van, D.L.: Contribution to Combinatorics on Words. Ph.D. thesis, Humboldt University, Berlin (1985)

A Study on the Modified Attribute Oriented Induction Algorithm of Mining the Multi-value Attribute Data

Shu-Meng Huang[1,2], Ping-Yu Hsu[1], and Wan-Chih Wang[3]

[1] Department of Business Administration, National Central University, Jhongli City,
Taoyuan County 32001, Taiwan (R.O.C)
[2] Department of Marketing and Distribution Management,
Hsing Wu Institute of Technology, LinKou District, New Taipei City 244, Taiwan (R.O.C)
[3] Graduate Institute of Management Sciences, Tamkang University, Danshui District,
New Taipei City 25137, Taiwan (R.O.C)
simon@mail.hwc.edu.tw

Abstract. Attribute Oriented Induction method （short for AOI） is one of the most important methods of data mining. The input value of AOI contains a relational data table and attribute-related concept hierarchies. The output is a general feature inducted by the related data. Though it is useful in searching for general feature with traditional AOI method, it only can mine the feature from the single-value attribute data. If the data is of multiple-value attribute, the traditional AOI method is not able to find general knowledge from the data. In addition, the AOI algorithm is based on the way of induction to establish the concept hierarchies. Different principles of classification or different category values produce different concept trees, therefore, affecting the inductive conclusion. Based on the issue, this paper proposes a modified AOI algorithm combined with a simplified Boolean bit Karnaugh map. It does not need to establish the concept tree. It can handle data of multi value and find out the general features implied within the attributes.

Keywords: Attribute Oriented Induction, Multi-Value-Attribute, Boolean bit, Karnaugh Map.

1 Introduction

Data mining extracts implicit, previously unknown and potentially useful information from databases. Many approaches have been proposed to extract information. According to the classification scheme proposed in recent surveys (Chen et al., 1996; Han and Kamber, 2001), one of the most important ones is the attribute-oriented induction (AOI) method. This approach was first introduced in Cai et al. (1990), Han etal. (1992), Han et al. (1993).

The AOI method was developed for knowledge discovery in relational databases. The input of the method includes a relation table and a set of concept trees (concept hierarchies) associated with the attributes (columns) of the table. The table stores the task-relevant data, and the concept trees represent the background knowledge. The core of the AOI method is on-line data generalization, which is performed by first

J.-S. Pan, S.-M. Chen, N.T. Nguyen (Eds.): ACIIDS 2012, Part I, LNAI 7196, pp. 348–358, 2012.

examining the data distribution for each attribute in the set of relevant data, calculating the corresponding abstraction level that the data in each attribute should be generalized to, and then replacing each data tuple with its corresponding generalized tuple. The major generalization techniques used in the process include attribute-removal, concept-tree climbing, attribute-threshold control, propagation of counts and other aggregate values, etc. Finally, the generalized data is expressed in the form of a generalized relation from which many kinds of rules can be discovered, such as characteristic rules and discrimination rules. For more details, please refer to the original papers (Cai et al., 1990; Han et al., 1992; Han et al., 1993).

Undoubtedly, the AOI method has achieved a great success. Because of its success, extensions have been proposed in the following directions: (1) extensions and applications based on the basic AOI approach (Han et al., 1993; Han et al., 1998; Lu et al., 1993), (2) more efficient methods of AOI (Carter and Hamilton, 1995; Carter and Hamilton, 1998; Cheung et al., 2000), (3) more general background knowledge (Hamilton et al., 1996; McClean et al., 2000), (4) integrating AOI with other information reduction methods (Hu and Cercone, 1996; Shan et al., 1995) and (5) proposing new variants of generalized rules (Tsumoto, 2000). (6) proposes a dynamic programming algorithm, based on AOI techniques, to find generalized knowledge from an ordered list of data(Chen and Shen, 2005).

AOI related algorithms conduct data induction with the help of concept hierarchies, which are needed for each inducted attribute and taken as a prerequisite to apply AOI. Concept hierarchies are the main characteristics of AOI and the primary reasons that AOI can conduct induction. However, the major characteristics have also become the major bumper of AOI applications. Two problems are rooted on the hierarchies. The first one is the scarce availability of creditable concept hierarchies. In many cases, users who need to summarize data for huge tables find the application of AOI unrealistic simply because the targeted attributes do not have sensible concept hierarchies. The second problem stems from the concept hierarchies are that concept hierarchies and associated attributes can only hold up to a single value. Unfortunately, many census data which would otherwise make very good applications of AOI, store data with multiple-value formats. For example, Table 1 shows a census data for several areas which are crime hot spots. The table contains the attributes of Area, Marital status, Gender and Education. Except Area, the other three attributes store data in set oriented multi-valued format. Each value in the set is a pair of <ordinal value, count>, where the ordinal value denotes a banded or categorical value which are totally ordered in their sorts; the count is the population in the area that are categorized into the corresponding group. For example, {<g1, 30> <g2, 70>}, in Area1 means that 30 persons in the area is has the gender of g1 and 70 of them has the gender of g2.

However, there are still ways AOI following a number of common shortcomings: (1) must have the concept trees, set up the concept trees varies from person to person, and finally summed up the rules will be different. In this paper, an algorithm is proposed to induct data organized in sets of ordinal and numeric pairs. Furthermore, no hierarchies are needed to induct the data. (2) can only deal with single-valued property. Unfortunately, a lot of information belong to many multi-value property values, such as census data, so the existing AOI response to the above questions, especially for multi-valued attribute value data, this paper presents a new AOI method, first the value of property 2 element of treatment, and then use Boolean function simplification, combined with Karnaugh Map Simplification of inductive methods, and finally re-scanned to identify a broad knowledge of statistics.

Table 1. 10 crime hot spots (DB)

Area	gender	age	education
1	<g1,30><g2,70>	<a1,20><a2,30><a3,50>	<e1,20><e2,10><e3,40><e4,30>
2	<g1,45><g2,55>	<a1,25><a2,35><a3,40>	<e1,15><e2,10><e3,35><e4,30>
3	<g1,65><g2,35>	<a1,35><a2,25><a3,40>	<e1,30><e2,40><e3,10><e4,20>
4	<g1,40><g2,60>	<a1,20><a2,40><a3,40>	<e1,10><e2,10><e3,40><e4,40>
5	<g1,35><g2,65>	<a1,30><a2,20><a3,50>	<e1,25><e2,5><e3,40><e4,30>
6	<g1,60><g2,40>	<a1,25><a2,25><a3,50>	<e1,20><e2,15><e3,35><e4,30>
7	<g1,20><g2,80>	<a1,10><a2,40><a3,50>	<e1,5><e2,30><e3,35><e4,30>
8	<g1,70><g2,30>	<a1,30><a2,40><a3,30>	<e1,10><e2,40><e3,40><e4,10>
9	<g1,40><g2,60>	<a1,20><a2,10><a3,70>	<e1,20><e2,20><e3,30><e4,30>
10	<g1,35><g2,65>	<a1,20><a2,30><a3,50>	<e1,20><e2,10><e3,40><e4,30>

g_1: the number of male, g_2:the number of female,

a_1: the amount of youth, a_2: the amount of adult, a_3: the amount of elder.

e_1: the number of primary school graduated, e_2: the number of high school graduated, e_3: the number of university graduated, e_4: the number of post graduated.

In view of these weaknesses, this paper proposed a new AOI method, In the algorithm, a translation procedure is deployed to translate each multi-value into a sequence of binary digits. The binary sequences are merged and hence inducted with Karnaugh map. The the last attribute value and then the same group of variables can be summarized in total by the following rules:

$$\{<g_1,L><g_2,H>\}\wedge\{<a_1,L><a_3,H>\}\wedge\{<e_1,L><e_3,H><e_4,H>\} \quad Y\%$$

That is:

Law and order problems will have the characteristics of Y% is the number of female more and more elderly and more than for tertiary education more.

2 The Data Structure and the Algorithm

This research aims to answer the limitation of multi-value attribute induction approach. Each multi-value attribute contains a set of pairs of ordinal and numeric values. The author proposed a method, which transforms each and numeric value into a Boolean bit and organize the bits into binary numbers according to the ordinal positions. The binary numbers are used to perform induction with the application of simplication Karnaugh Map.

2.1 Composition of Binary Values

The first step in the stage is to transform each numeric value in the set of ordinal and numeric value pair into a Boolean bit, which is equal to the result of comparing the corresponding number to the average of the numbers in the set. The second step then organizes the bit values into transformed binary numbers by ordering the Boolean bit according to corresponding ordinal position in the set.

Definition 1 (Data Transformation rules)

A. Let o be an ordinal number and n be a numeric value then <o, n> is an ordinal-numeric pair.

B. If S is a set of ordinal-numeric pairs, <o, n> in S and
 $b = (n >= \ (\sum_{<u,v>\in S} V)/|S|)$ then
 <o,b> is a ordinal-Boolean pair.

C. If $\{<o_i,b_i>\}$, where 1<= i <=n, is a set of n ordinal-Boolean pairs, then $b_1*2^{n-1} + b_2 *2^{n-2} \ldots+b_n$ is a transformed binary value.

For instance, in the first row of the education attribute in Table 1, the original value is {<e1,20>, <e2,10>, <e3,40>, <e4,30>}. According to Data Transformation rules, the corresponding set of ordinal-Boolean pairs is {<e1, 0>, <e2, 0>, <e3,1>, <e4,1>} and the transformed binary value is 0011. Table 2 shows the complete transformation of Table 1.

Table 2. The Booleans values of 10 crime hot spots (DB)

TID	gender	age	education
1	01	001	0011
2	01	011	0011
3	10	101	1100
4	01	011	0011
5	01	001	0011
6	10	001	0011
7	01	011	0111
8	10	010	0110
9	01	001	0011
10	01	001	0011

2.2 Binary Induction with the Application of Simplification Karnaugh Map

The induction process employs simplication Karnaugh map to summarize binary numbers without relying on induction hierarchies. Alan B. Marcovitz, (2005).

A Karnaugh map has a square for each '1' or '0' of a Boolean function. One variable Karnaugh map has 21 = 2 squares, Two variable Karnaugh map has 22 =4 squares, Three variable Karnaugh map has 23 = 8 squares, Four variable Karnaugh map has 24 = 16 squares etc. shown in Fig. 1.

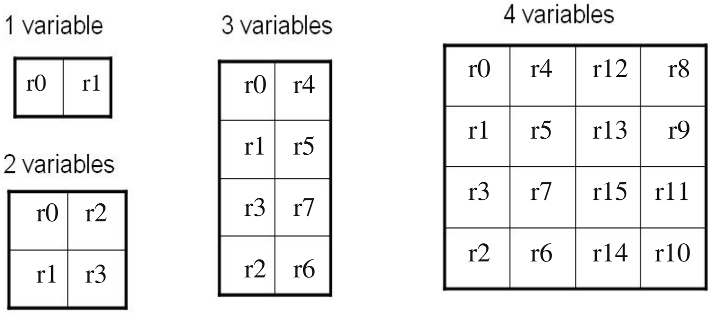


Fig. 1. Karnaugh map

The detailed processes have been illustrated by taking table 2 as an example. As shown previous, there are 4 variables in the ‘education’ table. The corresponded Karnaugh map is 16,(2^4), squares table. If we put the e_1, e_2, e_3, $e_{4,}$ as seen in figure 2, the counting of 0011 is 7 and 0101,0111,1100 is 1. The result of this transformation as illustrated in Fig. 2.

0	0	1	0
0	1	0	0
7	1	0	0
0	0	0	0

Fig. 2. The results of education attribute transferred to Karnaugh map

Based on the rule of Karnaugh map, we can take one variable out of the two adjacent variables. In Table 2, Education, there are two groups which variable is adjacent with each other, (r3, r5) and (r5, r7), and the variable r12 stands alone. The expression is indicating as shown follows.

$$\begin{aligned} F(e_1,e_2,e_3,e_4) &= \sum m(r3,r5,r7,r12) \\ &= (r3+r7)+r5+r12 \\ &= (0011+0111)+0101+1100 \\ &= 0\_11+0101+1100 \end{aligned} \tag{2}$$

Equation (2) reveals that the value ‘0011’ and ‘0111’ will be transferred to ‘0_11’.

Following similar procedure, the ‘age’ got 3 values whose Karnaugh map can be shown in Fig. 3.

0	0
5	1
3	0
1	0

Fig. 3. The results of age attribute transferred to Karnaugh map

After similar simplification process, the result of ‘age’ attribute can be described in equation (3).

$$\begin{aligned} F(a_1,a_2,a_3) &= \sum m(r1,r2,r3,r5) \\ &= (r1+r3)+r2+r5 \\ &= (001+011)+010+001 \\ &= 0\_1+010+001 \end{aligned} \tag{3}$$

Equation (3) reveals that the value ‘001’ and ‘011’ will be transferred to ‘0_1’.

It is not necessary to simplify ‘gender’ attribute, since there is only 2 values within it.

Table 3. The Boolean values of 10 crime hot spots (DB) has been converted into DB’

TID	gender	age	education
1	01	0_1	0_11
2	01	0_1	0_11
3	10	101	1100
4	01	0_1	0_11
5	01	0_1	0_11
6	10	0_1	0_11
7	01	0_1	0_11
8	10	010	0110
9	01	0_1	0_11
10	01	0_1	0_11

2.3 Stop Condition of the Induction

In this study, the definition of the “stop condition of the induction algorithm” is by way of re-checking the induction results in Table 3. Keep one data of the same attribute in the field in the table before re-checking, shown in Table 4. The inductive algorithm of attribute “education” and “age” can not be simplified anymore. Therefore, the algorithm must be stopped.

Table 4. The Boolean values of 10 crime hot spots (DB’)

TID	Gender	Age	Education
1	**01**	**0_1**	**0_11**
3	10	101	1100
6	10	0_1	0_11
8	10	010	0110

2.4 Inductive Rule

Sum up the values in the same column and row, 7 data are found to have the exact same value in each field. Then divide the above data by the total number of the data which gets an outcome of 70%. This is the inductive degree. The inductive rules are shown in Fig. 4. In this case, there is only one rule of high inductive degree. Due to the difficulty of interpreting binary values, the expression of the rule is then reduced to its symbolic form of raw data. As shown in Fig. 5, 0 means the value lower than the number of average threshold represented by L. 1 is the value higher than the number of average threshold represented by H. Where “baseline” indicates “do not care”, which means interpretation without transformation.

$$\text{rule 1 : } \{01\} \wedge \{0\ 1\} \wedge \{0\ 11\} \rightarrow 70\%$$

Fig. 4. Rule after induction

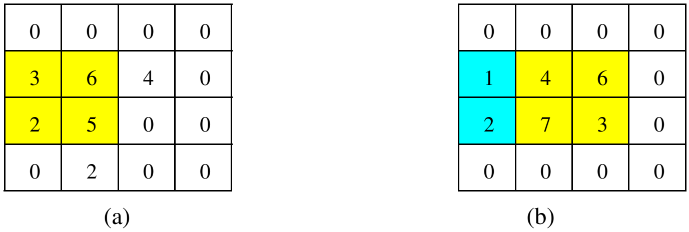
rule1 : {<g1,L><g2,H>} ^ {<a1,L><a3,H>} ^ {<e1,L><e3,H><e4,H>} → 70%

Fig. 5. Rule of easy-to-interpret after induction

2.5 Simplification Rules of the Modified Karnaugh Map

Simplification of Boolean algebra Karnaugh map in the most efficient, Karnaugh map provides a graphic (box) between the point of view of the relationship, to find two adjacent, 4, 8, 16 a group (re-select), you can simplify.

But the result has been repeated to select the option to repeat the question of attribution circle and lead to property value when summed up, the issue of double counting, it should be amended to do something. Circle option was to repeat, the majority of the combined selection and simplification to four variables Karnaugh map as an example, shown in Fi. 6.



0	0	0	0
3	6	4	0
2	5	0	0
0	2	0	0

(a)

0	0	0	0
1	4	6	0
2	7	3	0
0	0	0	0

(b)

Fig. 6. Modify Karnaugh map simplification rules

(1) Fig. 6 (a) based on the original Karnaugh map simplification rules, r1, r3, r5, r7 as a group, r5, r13 a, r7, r6 a, r5 and r7 of which are repeated to select, r1, r3, r5, r7 total number of (3 +2 +6 +5) = 16, r5, r13 Total Views (6 +4) = 10, r7, r6 total number of (5 +2) = 7, in order to avoid repeated induction calculation, so I chose to select r1, r3, r5, r7, chosen to give up r6, r13.

(2)Fig. 6 (b) based on the original Karnaugh map simplification rules, r1, r3, r5, r7 circle a group of selected, r5, r7, r13, r15 circle a group of selected, but r5, r7 was chosen to repeat, so I chose to select the number of larger (4 +7 +6 +3) = 20 of r5, r7, r13, r15 for a group, r1, r3 for the other group.

3 Performance Evaluation

3.1 Experimental Environment

In this section, a simulation is performed to empirically evaluate the performance of the proposed method. The KMAOI algorithm is implemented in C language and tested by using a PC with a P4 2.4G processor and 1024MB main memory under the Windows XP operating system. Since the AOI method has a much better time complexity than the KMAOI algorithm, the objective of the simulation is not to compare the number of times the PC runs, but to test the stability and inductive performance of the algorithm. This paper uses a real case of 2500 data of the MovieLens data sets. First, the meaningful attributes are filtered; the description of the attributes is shown in Table 5. Then the users' features on a rating of 4-5 points are inducted in zip code units from the questionnaire using the algorithm proposed, where the MovieLens data sets were collected by the GroupLens Research Project at the Minnesota University.

The data set consists of:
* 100,000 ratings (1-5) from 943 users on 1682 movies.
* Each user rates at least 20 movies.
* Simple demographic info of the users (age, gender, occupation, zip code).

The data was collected through the MovieLens web site (movielens.umn.edu) during a seven-month period from September 19th to April 22nd in 1998. The data has been treated which means that the users who have less than 20 ratings or do not have complete demographic information are removed from the data set.

Table 5. Description of attributes

Attribute	Number of Values	Values
Age	4	the amount of youth(under 20), adult(21~60), and elder(over 61)
Occupation	4	the number of engineer, educator, marketing and student
Movie_ type	4	the number of romance, comedy, thriller and action

3.2 Experimental Results and Performance Evaluation

(1) Induction Effect

After the actual algorithm under the circumstances of the absence from the hierarchy mechanism and hierarchy trees, the method proposed in this paper can induct the number of the data. The data of the tuples with the same value are voted respectively, in which the attribute data set of the user with movie rating of 4-5 points has be inducted. Then it is divided by the total number of the data. The induction degree is shown in table 6. The coverage of the induction summation of rule1, rule2 and rule 3 is 80.8%. It is obvious that the algorithm proposed has good induction ability. In the experiments, rule number 1 has the highest induction degree. It holds 1568 data out of 2500 for as high as 62.72 percent. The rule No.1 is listed, shown in Fig. 7, according to its original data format for a data interpretation.

Rule 1 : $\{<a_2, H><a_3, L>\} \wedge \{<o_1, H><o_4, L>\} \wedge \{<m_1, H><m_2, L>\} \rightarrow 62.72\%$

Fig. 7. Easy-to-read inducted rules

Table 6. Induction results of the real case

Age	Occupation	Movie_Type	Count	Percentage	Rule Number
_10	1__0	10__	1568	62.72%	(1)
_10	01__	10__	285	11.40%	(2)
_10	1__0	_10_	167	6.68%	(3)
_10	001_	10__	149	5.96%	(4)
_10	1__0	0_1_	124	4.96%	(5)
_10	_001	10__	98	3.92%	(6)
_10	01__	_10_	31	1.24%	(7)
_10	01__	0_1_	22	0.88%	(8)
_10	001_	0_1_	19	0.76%	(9)
_10	_001	_10_	16	0.64%	(10)
_10	001_	_10_	14	0.56%	(11)
_10	_001	0_1_	7	0.28%	(12)

Interpretation of the rule: among those users rating 4-5 points, 62.72% of them are between 21~60 years of age, and less are older than 61. Most of them are engineers in terms of the nature of work, while students are in the minority. They favor romantic movies more and comedic ones less.

(2) Stability

This paper is an empirical research. The data is divided into units with 300 data in each unit. The experiment is tested successively and accumulatively to investigate the stability of the inductive ability of the algorithm proposed by the study. The result shows that the algorithm has stable inductive ability no matter what size of the data is, as shown in Fig. 8.

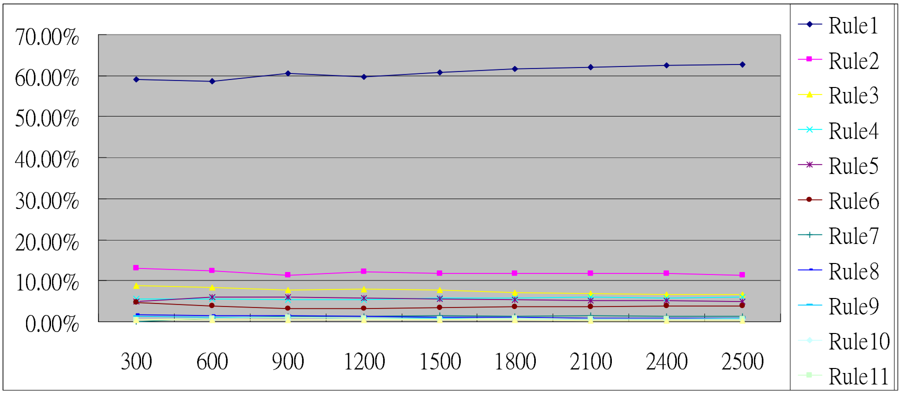


Fig. 8. Induction results of the real case

4 Conclusions

This paper proposes a modified AOI algorithm combining the simplified binary digits with Karnaugh Map. It is capable of dealing with data with multi-valued attributes without establishing the concept trees and extracting the general features implicit in the attributes.

This research concludes 3 contributions according to the empirical results. First, it solves the bottleneck problem of the traditional AOI method. There is no need to establish the concept hierarchies and concept trees during the inductive processes thus preventing from the heave workload. Second, the traditional AOI method is not able to deal with data with multi-valued attributes, while the method proposed by this paper can. Third, the data induction has very good inductive ability and stability.

References

1. Cai, Y., Cercone, N., Han, J.: An attribute-oriented approach for learning classification rules from relational databases. In: Proceedings of Sixth International Conference on Data Engineering, pp. 281–288 (1990)
2. Carter, C.L., Hamilton, H.J.: Performance evaluation of attribute-oriented algorithms for knowledge discovery from databases. In: Proceedings of Seventh International Conference on Tools with Artificial Intelligence, pp. 486–489 (1995)
3. Carter, C.L., Hamilton, H.J.: Efficient attribute-oriented generalization for knowledge discovery from large databases. IEEE Transactions on Knowledge and Data Engineering 10(2), 193–208 (1998)
4. Chen, M.S., Han, J., Yu, P.S.: Data mining: An overview from a database perspective. IEEE Transactions on Knowledge and Data Engineering 8(6), 866–883 (1996)
5. Cheung, D.W., Hwang, H.Y., Fu, A.W., Han, J.: Efficient rule-based attribute-oriented induction for data mining. Journal of Intelligent Information Systems 15(2), 175–200 (2000)
6. Hamilton, H.J., Hilderman, R.J., Cercone, N.: Attribute-oriented induction using domain generalization graphs. In: Proceedings of Eighth IEEE International Conference on Tools with Artificial Intelligence, pp. 246–252 (1996)
7. Han, J., Kamber, M.: Data Mining: Concepts and Techniques. Academic Press, New York (2001)
8. Han, J., Cai, Y., Cercone, N.: Knowledge discovery in databases: An attribute-oriented approach. In: Proceedings of International Conference on Very Large Data Bases, pp. 547–559 (1992)
9. Han, J., Cai, Y., Cercone, N.: Data-driven discovery of quantitative rules in relational databases. IEEE Transactions on Knowledge and Data Engineering 5(1), 29–40 (1993)
10. Han, J., Nishio, S., Kawano, H., Wang, W.: Generalization-based data mining in object-oriented databases using an object-cube model. Data and Knowledge Engineering 25, 55–97 (1998)
11. Lu, W., Han, J., Ooi, B.C.: Discovery of general knowledge in large spatial databases. In: Proceedings of 1993 Far East Workshop on Geographic Information Systems (FEGIS 1993), pp. 275–289 (1993)

12. Marcovitz, A.: Introduction to Logic Design. McGraw Hill, New York (2005)
13. Tokhenim, R.: Digital Electronics: Principles and Applications. McGraw Hill, New York (2005)
14. McClean, S., Scotney, B., Shapcott, M.: Incorporating domain knowledge into attribute-oriented data mining. International Journal of Intelligent Systems 15(6), 535–548 (2000)
15. Chen, Y.L., Shen, C.C.: Mining generalized knowledge from ordered data through attribute-oriented induction techniques. European Journal of Operational Research 166, 221–245 (2005)

A Hybrid Evolutionary Imperialist Competitive Algorithm (HEICA)

Fatemeh Ramezani[1], Shahriar Lotfi[2], and M.A. Soltani-Sarvestani[3]

[1] Computer Engineering Department, College of Nabi Akram, Iran
[2] Computer Science Department, University of Tabriz, Iran
[3] Computer Engineering Department, College of Nabi Akram, Iran
{f60_ramezani,soltani_mohammadamin}@yahoo.com,
shahriar_lotfi@tabrizu.ac.ir

Abstract. This paper proposes a new approach by combining the Evolutionary Algorithm and Imperialist Competitive Algorithm. This approach tries to capture several people involved in community development characteristic. People live in different type of communities: *Monarchy*, *Republic* and *Autocracy*. People dominion is different in each community. Research work has been undertaken to deal with curse of dimensionality and to improve the convergence speed and accuracy of the basic ICA and EA algorithms. Common benchmark functions and large scale global optimization have been used to compare HEICA with ICA, EA, PSO, ABC, SDENS and jDElsgo. HEICA indeed has established superiority over the basic algorithms with respect to set of functions considered and it can be employed to solve other global optimization problems, easily. The results show the efficiency and capabilities of the new hybrid algorithm in finding the optimum. Amazingly, its performance is about 85% better than others. The performance achieved is quite satisfactory and promising.

Keywords: Evolutionary Algorithm (EA), Imperialist Competitive Algorithm (ICA) and a Hybrid Evolutionary Imperialist Competitive Algorithm (HEICA).

1 Introduction

Recently there has been considerable amount of attention devoted to bio-inspiration and bio-mimicry, for solving computational problems and constructing intelligent systems. In the scope of computational intelligence it seems there are at least six main domains of intelligence in biological systems and wild life: Swarming, Communication and Collaboration, Reproduction and Colonization, Learning and Experience, Competition and Evolution [1].

Evolutionary algorithms, such as Genetic Algorithm (GA) [2], Simulated Annealing (SA), Particle Swarm Optimization (PSO) [3-4] and Ant Colony Optimization (ACO) [5] are computer simulation of natural processes such as natural evolution and annealing processes in materials.

The Imperialist Competitive Algorithm (ICA) was proposed by Atashpaz, in 2007 [6]. ICA is the mathematical model and the computer simulation of human social

J.-S. Pan, S.-M. Chen, N.T. Nguyen (Eds.): ACIIDS 2012, Part I, LNAI 7196, pp. 359–368, 2012.

evolution. It is one of the recent meta-heuristic algorithms based on a socio-politically, proposed to solve optimization problems [6]. ICA can be thought of as the social counterpart of GA because uses imperialism and imperialistic competition, socio-political evolution process, as source of inspiration.

People get together in organized groups or similar close aggregations. In each country people try to reach better position in society such as job, religious, economic and culture promotion. In other hand, it occurs among different countries. People who live together in country work with the leader to be better among other countries. Therefore it is a good idea to combine EA and ICA algorithms to increase performance. We use EA to imagery a point of view, person and competition between people, and ICA to imagery another point of view, country and competition between countries.

The paper is organized as follows. Section 2, provides a brief literature overview of the EA and ICA. In Section 3, some related work is presented. In Section 4, new approach and the motivation of the HEICA is presented. In Section 5, results are compared with other algorithms.

2 Background

As an alternative to the conventional mathematical approaches, the meta-heuristic optimization techniques have been widely utilized and improved to obtain engineering optimum design solutions. Many of these methods are created by the simulation of the natural processes. Evolutionary Algorithm aims to simulate natural selection with survival of fittest in mind. Imperialist Competitive Algorithm simulates the social political process of imperialism and imperialistic competition.

2.1 Imperialist Competitive Algorithm (ICA)

This algorithm starts by generating a set of candidate random solutions in the search space of the optimization problem. The generated random points are called the initial population (countries in the world). Countries are divided into two groups: imperialists and colonies. The more powerful imperialist, have greater number of colonies. The cost function of the optimization problem determines the power of each country. Based on their power, some of the best initial countries (the countries with the least cost function value), become Imperialists and start taking control of other countries (called colonies) and form the initial Empires [6].

Three main operators of this algorithm are Assimilation, Revolution and Competition. This algorithm uses the assimilation policy. Based on this policy the imperialists try to improve the economy, culture and political situations of their colonies. This policy makes the colony's enthusiasm toward the imperialists. Assimilation makes the colonies of each empire get closer to the imperialist state in the space of socio-political characteristics (optimization search space). Revolution brings about sudden random changes in the position of some of the countries in the search space. During assimilation and revolution a colony might reach a better position and has the chance to take the control of the entire empire and replace the current imperialist state of the empire.

In competition operator, imperialists attempt to achieve more colonies and the colonies start to move toward their imperialists. All the empires try to win and take possession of colonies of other empires. The power of an empire depends on the power of its imperialist and its colonies. In each step of the algorithm, all the empires have a chance to take control of one or more of the colonies of the weakest empire based on their power. Thus during the competition the powerful imperialists will be improved and the weak ones will be collapsed. After a while, the weaker empires will lose all their colonies and their imperialists will transform to the colonies of the other empires; at the end, all the weak empires will be collapsed and only one powerful empire will be left. All the colonies are randomly divided among the imperialists. More powerful imperialists take possession of more colonies [6].

Algorithm continues until a stop condition is satisfied such as just one imperialist will remain. In this stage the position of imperialist and its colonies will be the same.

2.2 Evolutionary Algorithm (EA)

There are many different variants of Evolutionary Algorithms (EAs). The common underlying idea behind all these methods is the same. The idea of EA is based on survival of fittest and it causes a rise in the fitness of the population in different generations. Based on the fitness function some of the better candidates are chosen, they are seed the next generation by applying recombination and mutation. Execution of these operators leads to a set of new candidates, the offspring. Replacement operator replaces new offspring in next generation, based on their fitness.

3 Related Works

Abderchiri and Meybodi [7] proposed two algorithms for Solving SAT problems: First, a new algorithm that combines ICA and LR. Secondly, a hybrid Hopfield network (HNN)-Imperialist Competitive Algorithm (ICA). The proposed algorithm (HNNICA) has a good performance for solving SAT problems.

Vahid Khorani, Farzad Razavi and Ahsan Ghoncheh [8] proposed R-ICA-GA (Recursive-ICA-GA) based on the combination of ICA and GA. A new method improves the convergence speed and accuracy of the optimization results. They run ICA and GA consecutively. Results show that a fast decrease occurs while the proposed algorithm switches from ICA to GA.

Jain and Nigam [9] proposed a hybrid approach by combining the evolutionary optimization based GA and socio-political process based colonial competitive algorithm (CCA). They used CCA–GA algorithm to tune a PID controller for a real time ball and beam system.

Razavi and others [10] studied the ability of evolutionary Imperialist Competitive Algorithm (ICA) to coordinate over current relays. The ICA was compared to the GA. The algorithms were compared in terms of the mean convergence speed, mean convergence time, convergence reliability, and the tolerance of convergence speed in obtaining the absolute optimum point.

4 A Hybrid Evolutionary Imperialist Competitive Algorithm

In this section, combines two algorithms to present a novel hybrid algorithm. The pseudo-code of the HEICA is presented in follow:

```
Procedure HEICA
Step 1: Initialization;
   Generate some random people;
   Randomly allocate remain people to others countries;
   Select more powerful leaders as the empires;
Step 2: Evolutionary Algorithm
   Roulette Wheel Selection;
   Crossover;
   Mutation;
   Replacement;
Step 3: Imperialist Competitive Algorithm
   People Assimilation; Move the people of each country
   toward their relevant leaders.
   People Revolutionary;
   Countries Assimilation; Move the leaders of each
   country toward their empires and move the people of
   each country as the same as their leaders.
   Countries Revolutionary;
   Imperialistic Competition; Pick the weakest country
   from the weakest empire and give it to the empire
   that has the most likelihood to possess it.
   Elimination; Eliminate the powerless empires.
Step 4: Terminating Criterion Control; Repeat Steps 2-3
   until a terminating criterion is satisfied.
```

4.1 Population

This algorithm starts by generating a set of candidate random solutions in the search space of the optimization problem. The generated random points are called the initial population which consists of persons. *Persons* in this algorithm are the counterpart of *Chromosomes* in GA and *Particles* in PSO which are array of candidate solutions. In human society, groups of people form community and are involved in community development.

4.2 Types of Community

There are different kinds of community:

- *Republic*: A republic is led by representatives of the voters. Each is individually chosen for a set period of time. This type community has president. Best of people select as candidate and each people vote to their president. The votes have been counted and candidate with highest elected.

- *Autocracy*: This type of community does not have leader. The people are free and there is no force. They do what they want.
- *Monarchy*: A monarchy has a king or queen, who sometimes has absolute power. Power is passed along through the family. This type of community has a monarch; People should follow her. The powerful person is selected as the monarch in each country. Different monarchy countries exist in empires; the best monarch of this type of country selects as empire.

4.3 Initialization

The algorithm starts with an initial random population called person. Some of the best person in the population selected to be the leaders and the rest form the people of these countries. Each country has an equal population.

The total power of a country depends on both the power of the leaders and the power of its people. This fact is modeled by defining the total power of a country as the power of the leader of the country plus a percentage of mean power of its people. The power of the people has an effect on the total power of that country:

$$T.P_{C_i} = Cost(Leader_i) + \xi mean\{Cost(People\ of\ country_i)\} \tag{1}$$

$T.P_{C_i}$ is the total power of i-th country.

Based on monarchy countries power, some of the best initial countries (the countries with the least cost function value), become Imperialists and start taking control of other countries and form the initial Empires. The best leaders of the countries determine the empire of the Empires. We select N_{Imp} of the most powerful countries to form the empires. To divide the countries among imperialists proportionally, we use the normalized cost of an imperialist by [6]:

$$N.C_{I_i} = C_{I_i} - \underbrace{Max}_{j}\{C_{I_j})\} \tag{2}$$

C_{I_i} is the cost of i-th imperialist and $N.C_{I_i}$ is its normalized cost. Having the normalized cost of all imperialists, the normalized power of each imperialist is defined by [6]:

$$N.P_{I_i} = \left| \frac{C_{I_i}}{\sum_{j=1}^{N_{Imp}} C_{I_j}} \right| \tag{3}$$

The normalized power of an imperialist is the portion of colonies that should be possessed by that imperialist. Then the initial number of countries of an empire will be [6]:

$$N.C_{I_i} = round(N.P_{I_i} * N_{Monarchy}) \tag{4}$$

$N.C_{I_i}$ is the initial number of countries of i-th empire and $N_{Monarchy}$ is the number of all monarchy countries. To divide the countries, for each imperialist we randomly choose $N.C_{I_i}$ of the monarchy countries and give them to it. These countries along with the imperialist will form i-th empire.

4.4 Assimilation and Revolution

After Evolutionary Algorithm operator accomplish, imperialist started to improve their countries and countries started to improve their people. HEICA has modeled these facts by moving all the countries toward the imperialist and all the people toward the leaders:

- *External*: External operations are among the countries. Assimilation occurs just among monarchy countries of each imperialist they move toward the empires. The country moves toward the imperialist it means all the people among these countries move in the same way, toward the empires. Just monarchy countries have an empire therefore assimilation is toward the empire. Revolution occurs in all countries. All the people of one country should move toward the same way because revolution is against the empire in monarchy countries. In other countries they try to improve their countries therefore they move with each other.
- *Internal*: Internal operations are among the people of the countries. Assimilation occurs in all countries they move toward the leaders. Revolution occurs in all countries, people try to get the position.

5 Evaluation and Experimental Results

For evaluating performance of the proposed algorithm, the simulation results are compared with results of EA, ICA, PSO and ABC [11]. *M* is mean of best results. Performance calculated from Equation 5 as follow for minimum optimization:

$$P_{Algorithm} = 100\% * (1 - \frac{M_{HEICA}}{M_{Algorithm}}) \qquad (5)$$

5.1 Benchmark Functions

Used common benchmark functions are listed in Table 1.The comparison results of F1-F10 are shown in Table 2. Best stability and convergence diagrams of different functions are shown in Fig. 1 and Fig. 2. In Table 3, the performance of HEICA algorithm is compared with PSO and ABC for high dimensional problem. In Table 4, the proposed HEICA algorithm was tested on benchmark functions provided by CEC2010 Special Session on Large Scale Global Optimization [12]. HEICA performs better than SDENS [13]. HEICA performs better on 1.2E5 and 6E5 FEs and jDElsgo[14] performs better on 3E6 FEs. To show the efficiency of HEICA in solving different function, logarithmic scale diagram is used. A logarithmic scale is a scale of measurement using the logarithm of a physical quantity instead of the quantity itself. Since, in this study, the values cover a wide range; logarithmic scale makes it easy to compare values.

5.2 Discussion

This paper proposes a novel hybrid approach consisting EA and ICA and its performance is evaluated using various test functions. Performance parameter shows HEICA perform better optimization than EA, ICA, PSO and ABC in all test functions. It shows that hybrid algorithm perform better optimization.

Table 1. Benchmark function (F1-F10)

Title	Function	Range
F1(Sphere)	$\sum_{i=1}^{D} x_i^2$	$-5.12 \leq x_i \leq 5.12$
F2(Rosenbrock)	$\sum_{i=1}^{D-1} 100(x_i^2 - x_{i+1})^2 + (1 - x_i)^2$	$-2.048 \leq x_i \leq 2.048$
F3(Rastrigin)	$\sum_{i=1}^{D} \left(x_i^2 - 10\cos(2\pi x_i) + 10\right)$	$-5.12 \leq x_i \leq 5.12$
F4(Griewangk)	$1 + \sum_{i=1}^{D} \left(\frac{x_i^2}{4000}\right) - \prod_{i=1}^{2} \left(\cos\left(\frac{x_i}{\sqrt{i}}\right)\right)$	$-600 \leq x_i \leq 600$
F5(Schwefel)	$\sum_{i=1}^{D} 418.9829 - x_i sin\left(\sqrt{\lvert x_i \rvert}\right)$	$-500 \leq x_i \leq 500$
F6	$\sum_{i=1}^{D} 10^{i-1} x_i^2$	$-10 \leq x_i \leq 10$
F7(Schaffer)	$0.5 + \frac{sin^2\sqrt{x^2 + y^2} - 0.5}{(1 + 0.001(x^2 + y^2))^2}$	$-100 \leq x, y \leq 100$
F8(Schwefel 1.2)	$\sum_{i=1}^{D} \left(\sum_{j=1}^{i} x_j\right)^2$	$-100 \leq x, y \leq 100$
F9(SumSquares)	$\sum_{i=1}^{D} i^2 x_i^2$	$-1 \leq x_i \leq 1$
F10(Ackley)	$-20e^{-0.2\sqrt{\frac{1}{D}\sum_{i=1}^{D} x_i^2}} - e^{\frac{1}{D}\sum_{i=1}^{D} \cos(2\pi x_i)} + 20 + e$	$-30 \leq x_i \leq 30$

6 Conclusion and Future Works

This paper proposed a novel hybrid approach consisting EA (Evolutionary Algorithm) and ICA (Imperialist Competitive Algorithm). The efficiency of HEICA is surveyed by comparing it evolutionary algorithm through a set of well-known multi-dimensional benchmark functions. The simulations indicate that the proposed algorithm has outstanding performance in speed of convergence and precision of the solution for global optimization, i.e. it has the capability to come up with non-differentiable objective functions with a multitude number of local optima in a reasonable time limit. We are working on using LEM (Learnable Evaluation Model) to improve results.

Table 2. The results achieved by HEICA on F1-F10

F	F1		F2		F3		F4		F5	
D.	10	50	10	30	10	50	10	50	10	50
Gen.	1000	2000	2000	4000	1000	2000	1000	3000	1000	2000
Pop.	150	500	250	600	150	400	450	750	150	300
EA	9.74E-03	5.15E+0	6.50E+00	7.42E+01	1.27E+00	7.81E+01	2.41E-02	1.06E-02	8.52E+00	1.58E+03
ICA	6.04E-04	8.18E-03	8.49E-01	1.11E+01	1.23E-01	2.46E+00	1.77E-03	2.15E-02	9.57E-01	9.59E+01
PSO	2.30E-35	2.14E-12	1.33E+00	3.32E+01	1.69E+00	6.72E+01	5.68E-02	1.00E-02	6.88E+02	7.76E+03
ABC	7.30E-17	1.14E-15	2.60E-03	5.79E+00	**0**	5.79-12	1.67E-17	**2.22E-17**	2.90E-02	3.19E+02
HEICA	**1.36E-39**	8.88E-23	**3.23E-03**	**7.06E+00**	**0**	**0**	**0**	3.33E-16	**1.27E-04**	**6.36E-04**
P_{EA}(%)	100	100	100	99	100	100	100	100	100	100
P_{PSO}(%)	100	100	100	97	100	100	100	100	100	100
P_{ABC}(%)	100	100	85	85	-	100	100	-823	100	100

F	F6		F7	F8		F9		F10	
D.	10	50	2	10	50	10	50	10	50
Gen.	1000	2500	1000	1000	1500	1000	2000	1000	2000
Pop.	150	450	150	150	200	150	500	150	200
EA	3.63E+05	1.55E+45	2.09E-02	1.09E+01	3.73E+04	1.01E-02	1.12E+02	1.52E+00	6.03E+00
ICA	1.78E+01	1.20E+28	1.34E-08	1.00E+00	1.19E+02	4.79E-04	9.53E-02	2.74E-01	7.15E-01
PSO	8.18E-16	2.16E+11	3.45E-04	-	-	2.61E-35	7.73E-11	4.26E-15	1.12E-03
ABC	6.32E-17	3.19E+02	6.88E-06	-	-	7.26E-17	1.57E-15	6.93E-15	2.02E-07
HEICA	**5.80E-37**	**1.11E-01**	**0**	**2.79E-39**	**9.06E-12**	**5.76E-39**	**8.77E-23**	**4.00E-15**	**6.51E-11**
P_{EA}(%)	100	100	100	100	100	100	100	100	100
P_{PSO}(%)	100	100	100	-	-	100	100	6	100
P_{ABC}(%)	100	100	100	-	-	100	100	42	100

Table 3. Comparing HEICA with ABC and PSO (high dimensions)

	D.	Pop.	Gen.	ABC	PSO	HEICA	P_{ABC}(%)	P_{PSO}(%)
F1	500	600	1500	2.26E+02	3.90E+03	**3.83E+01**	83	99
	1000	800	2000	1.46E+03	8.00E+03	**9.29E-01**	100	100
F2	500	600	1500	8.44E+03	1.99E+05	**4.53E+03**	46	98
	1000	800	2000	5.09E+04	4.21E+05	**4.03E+03**	92	99
F3	500	600	1500	1.93E+03	8.59E+03	**1.14E+03**	41	87
	1000	800	2000	6.05E+03	1.74E+04	**5.20E+02**	91	97
F4	500	600	1500	9.48E+02	5.10E+01	**1.23E+00**	100	98
	1000	800	2000	4.86E+03	2.98E+02	**5.10E+00**	100	98
F9	500	600	1500	5.23E+05	1.12E+07	**7.15E+04**	86	99
	1000	800	2000	1.84E+08	9.50E+07	**6.60E+03**	100	100
F10	500	600	1500	1.49E+01	2.09E+01	**1.88E+00**	87	91
	1000	800	2000	1.77E+01	1.74E+04	**4.89E+00**	72	100

Table 4. The results achieved by HEICA on the test suite

FEs	Algorithm	Shifted Elliptic	Shifted Rastrigin	Single-group Shifted and m-rotated Elliptic	Single-group Shifted m-dimensional Rosenbrock	$\frac{D}{2m}$-group Shifted and m-rotated Elliptic	$\frac{D}{2m}$-group Shifted and m-rotated Elliptic	$\frac{D}{2m}$-group Shifted and m-rotated Ackley	Shifted Schwefel
1.2E5	PSO	1.26E+10	2.36E+04	**4.06E+13**	**5.51E+08**	3.74E+10	4.31E+10	4.29E+02	9.92E+11
	SDENS[13]	5.01E+09	1.19E+04	5.10E+13	7.71E+08	1.56E+10	1.84E+10	**4.15E+02**	2.61E+11
	jDElsgo[14]	3.70E+09	1.09E+04	1.40E+14	2.39E+09	1.64E+10	2.32E+10	4.17E+02	7.99E+10
	HEICA	**9.76E+08**	**6.29E+03**	4.10E+13	8.07E+09	**3.54E+09**	**1.36E+10**	4.21E+02	**2.67E+10**
6E5	PSO	5.78E+08	2.36E+04	1.31E+13	1.19E+08	8.82E+09	1.52E+10	4.25E+02	1.00E+11
	SDENS	7.87E+06	7.09E+03	1.72E+13	7.41E+07	2.23E+09	5.14E+09	4.13E+02	2.69E+08
	jDElsgo	**8.99E+04**	3.95E+03	1.39E+13	6.82E+07	1.66E+09	4.10E+09	**2.99E+02**	1.01E+06
	HEICA	2.49E+06	**5.55E+02**	**8.42E+12**	**6.33E+07**	**9.55E+08**	**3.61E+09**	4.13E+02	**9.61E+05**
3E6	PSO	1.99E+07	2.36E+04	5.72E+12	7.16E+07	1.87E+09	6.45E+09	4.09E+02	8.55E+09
	SDENS	5.73e–06	2.21E+03	5.11E+12	5.12E+07	5.63E+08	1.88E+09	4.08E+02	**9.90E+02**
	jDElsgo	**8.86E-20**	**1.25E-01**	**8.06E+10**	**3.15E+06**	**3.11E+07**	**1.69E+08**	**1.44E+02**	1.53E+03
	HEICA	3.49E+04	2.62E+00	3.10E+12	5.24E+07	3.00E+08	1.09E+09	4.01E+02	3.73E+03

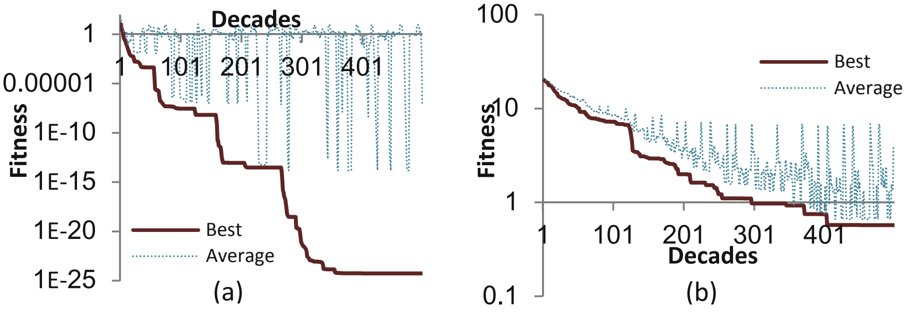


Fig. 1. Convergence diagram of HEICA: (a) F1 (D = 10) (b) F10 (D = 50)

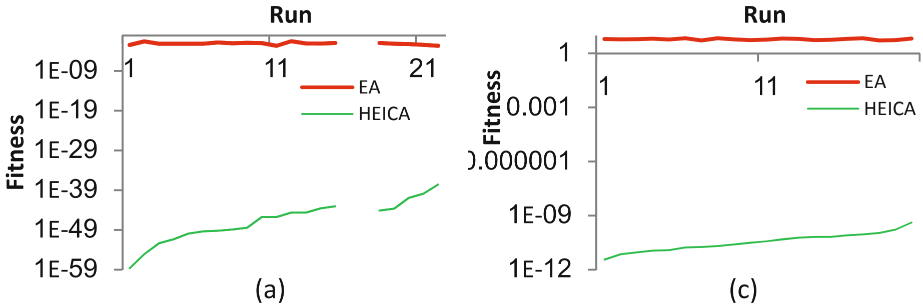


Fig. 2. Best stability diagram of HEICA: (a) F1 (D = 10) (b) F10 (D = 50)

References

1. Hajimirsadeghi, H., Lucas, C.: A Hybrid IWO/PSO Algorithm for Fast and Global Optimization. In: International IEEE Conference Devoted to 150 Anniversary of Alexander Popov (EUROCON 2009), Saint Petersburg, Russia, pp. 1964–1971 (2009)
2. Holland, J.H.: Adoption in Natural and Artificial Systems. University of Michigan, Ann Arbor (1975)
3. Kennedy, J., Eberhart, R.C.: Particle Swarm Optimization. In: Proceedings of IEEE International Conference on Neural Networks, Australia, pp. 1942–1948 (1995)
4. Eberhart, R., Kennedy, J.: A New Optimizer Using Particle Swarm Theory. In: Proceedings of the Sixth International Symposium on Micro Machine and Human Science (MHS 1995), pp. 39–43 (1995)
5. Dorigo, M., Maniezzo, V., Colorni, A.: Ant System: Optimization by a Colony of Cooperating Agents. IEEE Transactions on Systems, Man and Cybernetics, 1–13 (1996)
6. Atashpaz-Gargari, E., Lucas, C.: Imperialist Competitive Algorithm: An Algorithm for Optimization Inspired By Imperialistic Competition. In: Proceedings of the IEEE Congress on Evolutionary Computation, Singapore, pp. 4661–4667 (2007)
7. Abdechiri, M., Meybodi, M.R.: A Hybrid Hopfield Network-Imperialist Competitive Algorithm for Solving the SAT Problem. In: 3rd International Conference on Signal Acquisition and Processing (ICSAP 2011), pp. 37–41 (2011)
8. Khorani, A.V., Razavi, B.F., Ghoncheh, C.A.: A new Hybrid Evolutionary Algorithm Based on ICA and GA: Recursive-ICA-GA. In: The 2010 International Conference on Artificial Intelligence, pp. 131–140 (2010)
9. Jain, T., Nigam, M.J.: Synergy of Evolutionary Algorithm and Socio-Political Process for Global Optimization. Expert Systems with Applications 37, 3706–3713 (2010)
10. Razavi, F., Khorani, V., Ghoncheh, A., Abdollahi, H., Askarian, I.: Using Evolutionary Imperialist Competitive Algorithm (ICA) to Coordinate Overcurrent Relays. In: The 2011 International Conference on Genetic and Evolutionary Methods, GEM 2011 (2011)
11. Karaboga, D., Basturk, B.: Artificial Bee Colony (ABC) Optimization Algorithm for Solving Constrained Optimization Problems. In: Melin, P., Castillo, O., Aguilar, L.T., Kacprzyk, J., Pedrycz, W. (eds.) IFSA 2007. LNCS (LNAI), vol. 4529, pp. 789–798. Springer, Heidelberg (2007)
12. Tang, K., et al.: Benchmark Functions for the CEC 2010 Special Session and Competition on Large-Scale Global Optimization. s.l.: Nature Inspired Computation and Applications Laboratory, Technical report (2010), http://nical.ustc.edu.cn/cec10ss.php
13. Wang, H., Wu, Z., Rahnamayan, S., Jiang, D.: Sequential DE Enhanced by Neighborhood Search for Large Scale Global Optimization. In: WCCI 2010 IEEE World Congress on Computational Intelligence, pp. 4056–4062 (2010)
14. Brest, J., Zamuda, A., Fister, I., Maucec, M.S.: Large Scale Global Optimization using Self-adaptive Differential Evolution Algorithm. In: WCCI 2010 IEEE World Congress on Computational Intelligence, pp. 3097–3104 (2010)

Decoding Cognitive States from Brain fMRIs: The "Most Differentiating Voxels" Way

Chittibabu Namballa, Rahul Erai, and Krithika Venkataramani

Department of Computer Science and Engineering,
IIT Kanpur
Kanpur, India
{chitti,rahule,krithika}@cse.iitk.ac.in

Abstract. Since the early 1990s, fMRI has come to dominate the brain mapping field due to its relatively low invasiveness, absence of radiation exposure, and relatively wide availability. It is widely used to get a 3-D map of brain activity, with a spatial resolution of few milliseconds. We try to employ various machine learning techniques to decode the cognitive states of a person, based on his brain fMRIs. This is particularly challenging because of the complex nature of brain and numerous interdependencies in the brain activity. We trained multiple classifiers for decoding cognitive states and analyzed the results. We also introduced a technique for considerably reducing the large dimensions of the fMRI data, thereby increasing the classification accuracy. We have compared our results with current state-of-the-art implementations, and a significant improvement in the performance was observed. We got 90% accuracy, which is significantly better than the state-of-the-art implementation. We ran our algorithm on a heterogeneous dataset containing fMRI scans from multiple persons, and still got an accuracy of 83%, which is significant since it shows our classifiers were able to identify some basic abstract underlying neural activity, which are subject-independent, corresponding to the each cognitive states.

Keywords: machine learning, functional Magnetic Resonance Imaging, high dimensional data, Bayesian classifier, Support Vector Machine, nearest neighbor, brain image analysis.

1 Introduction

Since the introduction of brain fMRIs, study of human brain and its functioning, has received a tremendous boost. fMRIs were significantly better from its predecessors considering its high resolution, absence of brain radiation exposure and relatively wide availability. A single fMRI scan would give a brain image of about 15000 voxels in size. This rich data can be used for several studies including efforts to decode cognitive state of a person.

In this study, we used various machine learning tools to build several classifiers that can be used to decode subject's cognitive states in a particular time interval. Different stimuli can result the activation of different brain portions. i.e., the

J.-S. Pan, S.-M. Chen, N.T. Nguyen (Eds.): ACIIDS 2012, Part I, LNAI 7196, pp. 369–376, 2012.

brain regions getting activated while reading a book would be different from the regions that get activated while watching a movie. Once we learn this mappings, we can use them in several fields. These data can be potentially used to diagnose different medical conditions like Alzheimers disease.

1.1 Problem Statement

Theoretically, if a fMRI scan of sufficiently large resolution is given, one should be able to predict the exact cognitive state of the subject. However due to the practical constraints of computation, we restrict ourselves to a much simplified version of the problem. In this paper, we investigate the possibilities of using machine learning to determine whether the subject is reading a sentence or is seeing an image. Even in this restricted version, the problem is very challenging in a machine learning perspective. The fMRI data is extremely hyper dimensional, typically of the order of 1,00,000 dimensions, while we have only a very limited number of training samples at our disposal, that too highly noisy. So clearly, the accuracy of the classifier heavily depends on the dimensionality reduction and feature selection techniques that are used. We also describe a new method for identifying the most important features, reducing the dimensionality considerably, which in turn increased the accuracy of our system.

1.2 Functional Magnetic Resonance Imaging (fMRI)

In our study, we use Functional MRI or functional Magnetic Resonance Imaging (fMRI)[1] scans as an approximation for the actual neural activities of the subjects' brain. FMRI scanners measures the hemodynamic response (change in blood flow) related to neural activity in the the human brain . When neural cells are active they increase their consumption of energy from glucose and switch to less energetically effective, but more rapid anaerobic glycolysis. The local response to this energy utilization is to increase blood flow to regions of increased neural activity, which occurs after a delay of approximately 15 seconds. This hemodynamic response rises to a peak over 45 seconds, before falling back to baseline. This blood oxygen level dependent (BOLD) response is generally taken as an indicator of neural activity. The temporal response of the fMRI BOLD signal is smeared over several seconds. Given an impulse of neural activity, such as the activity in visual cortex in response to a flash of light, the fMRI BOLD response associated with this impulse of neural activity endures for many seconds.

2 Related Works

FMRI is a fairly new technology, developed in 1990s and is being heavily used in medical diagnostics. Little study has been done in using fMRI scans to train artificial intelligent agents to computationally model the human brain. Almost all the major work in the field of decoding cognitive states from brain fMRIs have been done in the past seven years, by Tom Mitchell and his colleagues at Carnegie Mellon University [2][3][4]. He and his team collected the star-plus dataset[5], which is widely used in studies involving fMRI data. We also used this dataset for training/testing of our classifiers, and used Mitchell's results as a benchmark to evaluate our algorithm. Our work is based on the paper by Mitchell et al.[2], in which they investigated the possibility of building computational models, that can be used to predict whether a person is seeing a picture or reading a sentence given his/her brain fMRI scan. They used a variety of dimensionality reduction techniques based on some simple heuristics, and trained a series of classifiers including Naive Gaussian Classifier, Support Vector Machines and Nearest Neighbour classifier. Despite of the hyper dimensionality of the data, and relatively basic classification techniques, they managed to get impressive results, considering how complex and interdependent the brain functions really are. They also trained cross-subject classifiers and still got decent performance.

Apart from Mitchell et al., there has been some efforts in employing machine learning techniques on data collected using other similar devices that records brain activity. For example, Blankertz et al. trained classifiers on EEG data collected from a single subject [6]. Friston et al. had dealt with an exact opposite problem than ours, where they tried to predict the future values of a voxel based on its previous states.

The work we present here are based on Mitchell's works[2][3][4], augmented with a new technique to reduce the dimensionality resulting in improved accuracy.

3 Dataset and Data Collection Procedure

We used Carnegie Mellon University's star-plus fMRI dataset[5] for our experiments. This data consists of 6 subjects and about 54 trials per subject. In each trail, the subject will be shown an image followed by a sentence (or vice versa) and he/she was asked to decide whether the sentence given accurately describes the image shown. Meanwhile, brain activities of the subject were continuously monitored and recorded using fMRIs.

3.1 Procedure

As explained before, the experiment consists of a set of trials, and the data was partitioned into trials. For some of these intervals, the subject simply rested, or gazed at a fixation point on the screen. For other trials, the subject was shown a picture and a sentence, and instructed to press a button to indicate whether the sentence correctly described the picture. For these trials, the sentence and picture were presented in sequence, with the picture presented first on half of the trials, and the sentence presented first on the other half of the trials. Forty such trials were available for each subject. The timing within each such trial is as follows:

- The first stimulus (sentence or picture) was presented at the beginning of the trail (1st image).

– Four seconds later (9th image) the stimulus was removed, replaced by a blank screen.
– Four seconds later (17th image) the second stimulus was presented. This remained on the screen for four seconds, or until the subject pressed the mouse button, whichever came first.
– A rest period of 15 seconds (30 images) was added after the second stimulus was removed from the screen. Thus, each trial lasted a total of approximately 27 seconds (approximately 54 images). For further information, please refer [5].

4 Our Approach

Every classification problem has two phases: Extracting the useful features, and training classifiers to do the actual classification. The first phase, i.e. feature selection/dimensionality reduction has a very important role in this particular project, as we are dealing with data in extremely high dimensions(150000), prone to noise. Hence, the accuracy of the classifier is heavily dependent on the method one uses to reduce the dimension of the data. This is the area where our approach differs from that of T. M. Mitchell and his team. We were able to come up with a better algorithm that reduced the dimension of the data considerably, without sacrificing the separability of the data distribution.

4.1 Dimensionality Reduction/Feature Selection

We used a two level mechanism to reduce the dimension. While the first step reduces the effect of noisy voxels, second step chooses voxels which are best suited to distinguish between sentences and pictures.

Step 1: Choosing Most Active Voxels. In first level, we adapted Mitchell's method[2] to choose the most active n voxels. This was mainly to reduce the noise in the voxel data. We divided the fMRI data into three different classes corresponding to fixation intervals, intervals when an image was shown, and intervals when a sentence was given. Since during the fixation time, subject was asked not to think about anything and to stare a blank screen, we considered all the voxel activities in this interval as background noise. Then we constructed a distribution per voxel per class (classes being sentence and picture), and did a t-test on each of this distribution to the background noise distribution, and choose n voxels which gave maximum t-test values. Since t-test values gives us the amount by which the two distributions vary from one another, this gave us voxels whose values vary considerably when an image or sentence is shown, which can be considered as most active voxels.

Step 2: Choosing Most Differentiating Voxels. Even though now we have a set of most active voxels from step 1, most of the voxels out of them won't be that useful in distiguisihing *pictures* from *sentences*. In other words, some voxels would respond equally to both pixels and voxles, making them an *active* voxel, but still a less attractive option to distinguish *pictures* from *sentences*.So, in this step, discard these voxels, further reducing the number of voxels selected by step 1. Here also we use a variant of t-testing. As the first step, record the values of each voxel when a picture is shown and a sentence is given to read. Now for each voxel, we again created two distribution, one corresponding to its activity when an image is shown and another corresponding to its activity when a sentence was given.

The difference between the two distribution of a voxel shows how efficient is that voxel in differentiating pictures and sentences. If the intersected area is larger, that means that both the distributions are more similar, making that voxel less suitable for differentiating pictures and sentences. On other hand, if the common area is small, that means, that voxel exhibit different response while seeing a picture than when reading a sentence, making it an excellent choice to be used in classification. Now, assuming that the distributions follow normal distribution, the intersected area is directly proportional to their variances and inversely proportional to the difference of the means of the distributions. So our objective is to discard those voxels whose t values are very small, where,

$$t = \frac{|\mu_1 - \mu_2|}{\sigma_1 + \sigma_2} \tag{1}$$

So, after this step, we are left with n voxels, which are best suited to differentiate between pictures and sentences.

4.2 Classification

To do the actual classification, we built three classifiers Gaussian Naive Bayesian (GNB), Support Vector Machines (SVM) and Nearest Neighbor classifier(KNN) [7] and trained them using the reduced dimension data obtained by the previous steps.

To train a support vector machine, we used linear kernel support vector machine. And for the K- nearest neighbor classifier, we considered Euclidean distance as a distance metric with k=1, 3,5,7 and 9.

5 Experimental Results

We implemented the system using MATLAB 7.10. and compared previously published results[2][3][4] with the results obtained by our approach. We found that our method significantly improves the accuracy over the existing methods.

5.1 Comparison between Mitchell et al.'s Approach vs. Our Approach

For comparing with Mitchell et. al.'s approach, we trained different classifiers, one for each subject and took the average accuracy over all the subjects. We experimented by considering different number of active voxels and chose the one which gave the best results.

Table 1. The table shows that augmenting our feature selection method improved the accuracies of *all* the classifiers that we trained for

Classifier	Mitchell et. al's results [2]	Our Results
GNB	82	86
SVM	89	90
1NN	78	84
3NN	82	84
5NN	82	82

In the above table, there is a comparison between the the results obtained by Mitchell et al's method, with the results that we got after enhancing the algorithm with our new dimensionality reduction technique. From the above presented data, we can easily observe that our approach is giving significantly better performance than their approach.

5.2 Effect of Number of Active Voxels on Accuracy of the Classifier

As we noted earlier, the number of voxels that we consider after feature selection affects the performance of the classifier. The graph given in Fig. 1 shows how the accuracy of classifiers varies with respect to the number of voxels used for classification.

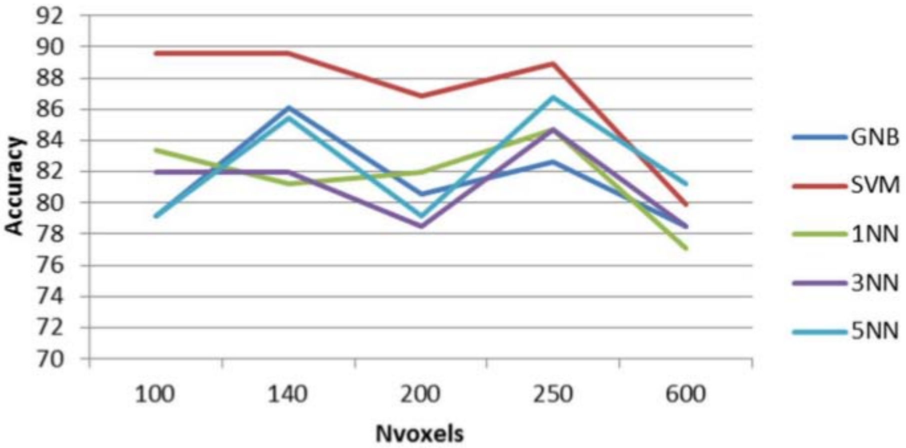


Fig. 1. *Accuracy* vs *Number of features*: Our observations shows that limiting the number of features to 250 gives the best results for all the classifiers

5.3 Cross-subjects Classification Results

In the previous section, we compared our results with Mitchell et. al.s[2] results by taking the average of all the single-subject classifiers. Apart from that, we also tried to build a cross-subject classifier. The idea was to train a single model using multiple subject's fMRI data, and test the trained model using a new subject's dataset. However, there were some practical problems associated with this approach. Since we don't have a unified spacial coordinate system to identify the voxels, it would be extremely difficult to identify which one of the voxels in subject A corresponds to a given voxel in subject B. Due to the huge resolution of the fMRI scans, it's almost impossible to calibrate fMRI scanners to produce scans in a unique coordinate system.

Table 2. Cross-subjects classification results

Classifier	Accuracy
GNB	77.08
SVM	79.86
1NN	81.94
3NN	77.08
5NN	82.64

One workaround for this problem would be, to mix all the subject's data together, and then randomly partition them into testing and training sets, so that both the sets would hopefully contain data from all the subjects. Even though this doesn't solve the problem, since both the test and train set contain data from all the subjects, slight spacial inconsistencies of the voxels would get averaged out. This is the closest that we can get to a cross-subject testing. Even with this heuristics, we were able to get impressive results using our algorithm, on cross-subject datasets. Table 2 summarizes the results we observed.

6 Conclusion

In this paper, we investigated how well machine learning techniques can be used to decode cognitive states from brain fMRIs. We studied the past researches in the field, and were able to augment the existing system with a new feature selection technique, which was able to identify the most relevant voxels that can be used to distinguish the given cognitive states.We built three different classifiers (GNB, SVM, KNN) and compared the results with the current state of the art systems results. The best accuracy that we got for the single-subject dataset was of 90%, using SVMs. Our algorithm showed surprisingly accurate performance for multi-subject dataset, implying that the feature selection algorithm that we used, indeed implicitly managed to identify the parts of the brain, which are accountable for the given cognitive states.

Since our algorithm is quite general in nature, it can be used to differentiate between any two cognitive states. For example, it can be used over the star-plus dataset to analyze "ambiguity" in the sentences. In future,apart from detecting the cognitive state, this study can be generalized in identifying parts of brain that are responsible for specific cognitive states.

Acknowledgment. We would like to thank Tom Mitchell et al, at CMU for providing the dataset, some useful functions and an excellent documentation of the dataset. We also would like to thank the Department of Computer Science, IIT Kanpur for providing the resources and support for our study.

References

1. Functional magnetic resonance imaging, http://en.wikipedia.org/wiki/Functional_magnetic_resonance_imaging
2. Mitchell, T.M., Hutchinson, R., Niculescu, R.S., Pereira, F., Wang, X., Just, M., Newman, S.: Learning to decode cognitive states from brain images. Machine Learning 57(1), 145–175 (2004)
3. Wang, X., Hutchinson, R., Mitchell, T.M.: Training fmri classifiers to detect cognitive states across multiple human subjects. In: Proceedings of the 2003 Conference on Neural Information Processing Systems, Citeseer (2003)
4. Wang, X., Mitchell, T.M., Hutchinson, R.: Using machine learning to detect cognitive states across multiple subjects. CALD KDD Project Paper (2003)
5. Cmu's starplus fmri dataset archive, http://www-2.cs.cmu.edu/afs/cs.cmu.edu/project/theo-81/www/
6. Blankertz, B., Curio, G., Mller, K.R.: Classifying single trial eeg: Towards brain computer interfacing. advances in neural inf. Proc. Systems 14(1), 157–164 (2002)
7. Duda, R.O., Hart, P.E., Stork, D.G., et al.: Pattern classification, vol. 2. Wiley, New York (2001)

A Semi-Supervised Method for Discriminative Motif Finding and Its Application to Hepatitis C Virus Study

Thi Nhan Le and Tu Bao Ho

School of Knowledge Science
Japan Advanced Institute of Science and Technology
1-1 Asahidai, Nomi City, Ishikawa, 923-1292 Japan

Abstract. Finding discriminative motifs has recently received much attention in biomedical field as such motifs allows us to characterize in distinguishing two different classes of sequences. Although the developed methods function on labeled data, it is common in biomedical applications that the quantity of labeled sequences is limited while a large number of unlabeled sequences is usually available. To overcome this obstacle, this paper presents a proposed semi-supervised learning method that enables the user to exploit unlabeled sequences to enlarge labeled sequence set, leading to improvement of the performance in finding discriminative motifs. The comparative experimental evaluation of the proposed semi-supervised learning shows that it can improve considerably the predictive accuracy of the found motifs.

Keywords: Discriminative motif finding, semi-supervised learning, self-training technique, hepatitis C virus.

1 Introduction

A *sequence motif* is generally understood as a nucleotide or amino acid pattern that is widespread and has a biological significance, commonly as transcription factor binding sites (TFBSs). Traditionally, the motif finding problem has been dominated by generative models using only one sequence class to produce descriptive motifs of the class. Recently, research focuses on discovering of discriminative motifs those can be used to distinguish sequences belonging to two different classes, typically methods using hidden Markov model (HMM) [1], probabilistic models [2], association mining with domain knowledge [3].

It is common in biomedical applications that the number of labeled sequences is usually small while a large number of unlabeled sequences is available. Take the case of our hepatitis study. It is well known that using interferon combined with ribavirin (IFN/RBV) is currently the standard therapy of hepatitis C virus (HCV). However, only near a half of the HCV infected individuals achieves sustained viral response (SVR) by this therapy, and the genetic basis of resistance to antiviral therapy remains unknown [4]. The non-structure protein 5A

J.-S. Pan, S.-M. Chen, N.T. Nguyen (Eds.): ACIIDS 2012, Part I, LNAI 7196, pp. 377–384, 2012.

(NS5A protein) in HCV genome is known as the protein most reported to be implicated in the interferon resistance [5]. However, gathering the biggest resource from Los Amalos HCV database [6], we can only have 134 NS5A non-SVR sequences and 93 SVR sequences to IFN/RBV therapy. In addition, from Genbank [7] and HVDB [8], we obtained about 5000 NS5A sequences belonging to genotypes 1a, 1b, 1c, 2a and 2a. In [9], a method for discovering discriminative motifs based on separate-and-conquer strategy that allows us to be able to detect the DMOPS (discriminative multiple occurrences per sequence) motifs when the number of labeled sequences is small. When applying the algorithm to the NS5A data, the results is promising but the predictive accuracy is still less than 70% on the testing data.

This work aims to enrich the small labeled dataset from a large unlabeled dataset in order to improve the average accuracy of discriminative discovered motifs. Our framework use the sequence matching mechanism on unlabeled data to create new training dataset, and the sequences that contain the most discriminative motif are added in each iteration. These selected sequences with predicted label are able to help improving the sufficiently and diversity of training dataset, from then help to find more discriminative motifs.

2 Background

2.1 Discriminative Motif Discovery

Protein motif is particular amino acid sequence that is characteristic of a specific structure or function of the molecule, and thus it appears more frequently than it is expected. The identification of these motifs from protein sequences plays an important role in controlling the cellular localization, can highlight interactive regions shared among proteins, or regions that are characteristic of a specific function/structure of molecule [3].

Traditionally, motif finding problem has been dominated by generative models using only one sequence class to produce descriptive motifs of the class. Recently, many studies focus on discovering of discriminative motifs that can be used to distinguish sequences belonging to two different classes [1–3, 9–11]. Discriminative motifs are those occurring more frequently in one set of sequences and not occur in another set. These motifs can help to classify well a sequence into certain class or to describe the characteristics of a class.

DMOPS is one of motif models categorized by [2] based on counting the number of total occurrences of motif in sequences. Because the significant regions in sequence are generally better preserved, our previous work focused on DMOPS motifs to detect the relationship between HCV NS5A protein and IFN/RBV therapy effect. These DMOPS are promising as they present many patterns that were not known previously.

2.2 DMOPS Motif Finding

In this part, we summarize the main ideas of our DMOPS motif discovery method, as well as key concepts used in [9].

Given two sets of positive and negative sequences, we find a minimal set of DMOPS motifs satisfying two conditions: (1) *Complete*: each sequence in positive class, denoted by Pos, contains at least one found motif, (2) *Consistent*: found motifs occur in sequences of positive class but do not occur in sequences in negative class, denoted by Neg.

A subsequence will be a DMOPS motif when it satisfies both α-*coverage* and β-*discriminant* thresholds. Given parameters α $(0 < \alpha < 1)$ and β $(0 < \beta < 1)$ a subsequence P is an α-*coverage* for class positive if

$$\frac{|cover_{Pos}(P)|}{|Pos|} \geq \alpha,$$

and is a β-*discriminant* for class positive if

$$\frac{|cover_{Pos}(P)|}{|cover_S(P)|} \geq \beta,$$

where $cover_{Pos}(P)$ is the set of sequences in class positive that each sequence contains a given subsequence P and $cover_S(P) = cover_{Pos}(P) \cup cover_{Neg}(P)$. If P is both α-coverage and β-discriminant for class positive, we will say P is α, β-strong for class positive. Similar concepts can be defined for class negative.

Because the number of sequences of class positive and negative is small in [9], DMOPS motifs that are found from training set are not enough to divide the dataset separately, and hence, the average accuracy of assessment on test dataset is achieved 68.8%. This result suggests us to use unlabeled sequences to search more general DMOPS motifs that can help us distinguish well two sets of positive and negative sequences and improve the predictive accuracy.

In the next section, we will present our semi-supervised learning framework and the method to enlarge the number of sequences for small labeled dataset.

3 Semi-Supervised Discriminative Motif Finding

Self-training is one of the most commonly used technique for semi-supervised learning [12]. In self-training, a classifier is first trained with the small amount of labeled data. The classifier is then used to classify the unlabeled data. Only the unlabeled data with their predicted labels having the most confident score are added to the training dataset. After that, the classifier is re-trained and this procedure is repeated several times until the convergence is reached.

We develop a semi-supervised learning (SSL) framework based on the idea of self-training technique to enlarge gradually the core DMOPS motif set. This framework works under the assumption: if sequences contain the same DMOPS motifs, they are likely to be of the same class. In other word, the more an unlabeled sequence contains DMOPS motifs, the more this sequence belongs to the class of those DMOPS motifs.

3.1 Our Framework

The framework (Figure 1) works as followings: First, from a small set of labeled sequences, DMOPS motif algorithm searches for discriminative motifs including motifs in class positive and negative. Next, these motifs will be used as a core set of discriminative motifs to match with unlabeled sequences. When a matching happens, a rank of an unlabeled sequence will be calculated. After that, labels are assigned to unlabeled sequences based on their rank. Finally, the unlabeled sequences with predicted label will be candidates and added to training dataset for the next iteration. This procedure will be repeated until (1) the set of sequence candidates is stable or (2) the maximum number of iterations is achieved.

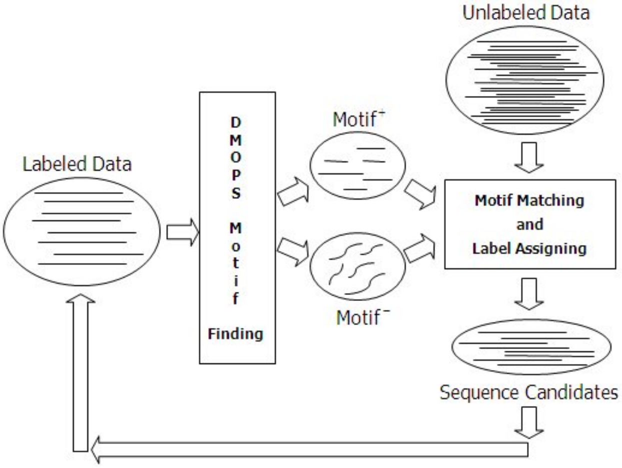


Fig. 1. A framework of semi-supervised discriminative motif finding

3.2 DMOPS Motif Finding Algorithm

The DMOPS motif finding algorithm is described in Algorithm 1. Given two sets of positive sequences *Pos* and negative sequences *Neg*, algorithm 1 will find a minimal set of DMOPS motifs satisfying condition *Complete* and *Consistent*. In this algorithm, $Motif(Pos, Neg, UL, \alpha, \beta, \gamma)$ is an exhaustive search procedure that expands a subsequence one position to the left or to the right, starting with length's subsequence is 1.

3.3 Motif Matching and Label Assigning Algorithm

We propose a discriminative motif matching method to perform the label assignment for unlabeled sequences based on the argument: discriminative motifs are properties of a class, and hence, if a sequence fills with discriminative motifs, it will likely belong to class of these motifs. During matching discriminative motifs and unlabeled sequences, we have to reject sequences containing both motifs in class positive and negative to guarantee the consistent condition of DMOPS motif. In addition to, ranking unlabeled sequences according to the number of DMOPS motifs that they contain and choosing the highest rank to assign label can help to confirm that if an unlabeled sequence contains many DMOPS motifs of a class, its chance to be a member of that class is high.

Algorithm 1. DMOPS motif finding algorithm [9].

```
**DMOPS Motif** $(Pos, Neg, mina, minb, \gamma)$

$MotifSet = \phi$
$\alpha, \beta \leftarrow$ **Initialize**$(Pos, mina, minb)$
**while** $Pos \neq \phi$ & $(\alpha, \beta) \neq (mina, minb)$ **do**
    $NewMotif \leftarrow$ **Motif**$(Pos, Neg, UL, \alpha, \beta, \gamma)$
    **if** $NewMotif \neq \phi$ **then**
        $Pos \leftarrow Pos \setminus Cover^{+}(NewMotif)$
        $MotifSet \leftarrow MotifSet \cup NewMotif$
    **else**
        **Reduce**$(\alpha, \beta)$
    **end if**
    $MotifSet \leftarrow$ **PostProcess**$(MotifSet)$
**end while**
return$(MotifSet)$
```

With a motif in class positive or negative, denoted by $motif^{+}$ or $motif^{-}$, a set of unlabeled sequences denoted by U is divided into two parts, one part contains the motif, the other part does not contain, denoted by $U1^{+}/U2^{+}$ or $U1^{-}/U2^{-}$. With the purpose to search more DMOPS motifs, we perform union of $U1_i^{+}/U1_i^{-}$ corresponding to $motif_i^{+}/motif_i^{-}$, where $i = 1..N$, N is number of positive/negative motifs in core motif set.

The motif matching and label assigning algorithm is described in Algorithm 2. In this algorithm, $checkConsistency(U1^{+}, U1^{-})$ is a procedure which performs a consistent condition check and $chooseHighestRank(U1^{+}, U1^{-})$ performs our SSL's assumption.

4 Application to Study NS5A Region of Hepatitis C Virus and Resistance to IFN/RBV Therapy

4.1 Data

In this study, all sequences are NS5A region of HCV genotype 1b that each sequence contains 447 amino acids. We used two kind of datasets as follows:

- *Labeled dataset*: including 28 sequences SVR and 49 sequences non-SVR from the Los Amalos HCV database [6].
- *Unlabeled dataset*: including 1437 sequences from HVDB [8] and 168 sequences from GenBank [7].

4.2 Experiment

This experiment focuses on validating and comparing the accuracy assessment of DMOPS motif finding algorithm before and after enlarging labeled dataset. Therefore we perform 3-fold cross validation five times with parameters are set as follows:

Algorithm 2. Motif matching and label assigning

```
Input: $motif^+$, $motif^-$, U
Output: $U1^+$, $U1^-$

$U1^+ = \phi$, $U1^- = \phi$
for each $motif \in motif^+/motif^-$ do
    if $motif$ match $sequence \in U$ then
        if $sequence \in U1^+/U1^-$ then
            increase $rank(sequence)$ by 1
        else
            $U1^+/U1^- \leftarrow sequence$
            $rank(sequence) \leftarrow 1$
        end if
    end if
end for
$checkConsistency(U1^+, U1^-)$
$chooseHighestRank(U1^+, U1^-)$
return$(U1^+, U1^-)$
```

In DMOPS motif finding experiments, the parameters $minAlpha$, $alphaStep$, $minBeta$, $betaStep$ and $maxLength$ are set to 0.1, 0.05, 0.6, 0,05 and 4, which are identified in [9] as their most appropriate values, respectively. These parameters are fixed during the iteration process of SSL.

In motif matching and label assigning experiments, 1437 unlabeled sequences are used and repeated for each iteration to pick sequence candidates out. The maximum number of iterations are set to 10 and the parameter to choose the highest rank of sequence is 1. Because the number of sequences in training set is small, we consider the case of one match between DMOPS motif and unlabeled sequence that is enough to assign label for that unlabeled sequence.

Table 1 shows the experiment results of comparing the accuracy of DMOPS motif finding algorithm before and after performing SSL (about 5% of accuracy is increased). The second and third columns are the average accuracy of assessment in testing data. $Agl1$ is DMOPS motif finding algorithm and $Alg1 + Alg2$ is DMOPS motif finding algorithm and SSL. In five times of 3-fold cross validation, the accuracy of $Alg1 + Alg2$ is less than the accuracy of $Agl1$ in one time; and four remaining times, the accuracy of $Alg1 + Alg2$ is larger than 70% and larger than the accuracy of $Agl1$. This can be explained by the number of DMOPS motifs that is found during SSL process. In addition, the appearance of new motifs makes iteration happen more times, leading to DMOPS motifs are more precise. The accuracy of $Alg1 + Alg2$ varies quite different because of starting from the small number of labeled data.

Table 1. Accuracy of DMOPS algorithm before and after performing SSL

Algorithm	Alg1	Alg1+Alg2
The 1^{st} 3-fold	0.77	0.79
The 2^{nd} 3-fold	0.69	0.91
The 3^{rd} 3-fold	0.64	0.73
The 4^{th} 3-fold	0.67	0.48
The 5^{th} 3-fold	0.71	0.84
	0.696	0.75

5 Conclusion

We have explored the use of self-training-based semi-supervised learning to enlarge training set of discriminative motif finding problem in case the number of labeled data is small. The proposed method works in an iterative procedure to choose the best match sequence candidates among unlabeled sequences. Experiment results show that with more data for training set, the DMOPS motif finding algorithm can obtain higher accuracy.

In this work, our method was developed based on assumption of containing the same discriminative motif among sequences. However, there are many cases that a sequence probably does not contain a discriminative motif but it may belong to the class of motif. There is reason to think that this sequence has a similarity in some respect with the sequence containing discriminative motif, for example gene distance among sequences.

Acknowledgments. This work is supported by JSPS's project No. 19300045. We would like to thank Ngoc Tu Le for providing source code of DMOPS motif finding algorithm.

References

[1] Lin, T., Murphy, R.F., Bar-Joseph, Z.: Discriminative motif finding for predicting protein subcellular localization. IEEE/ACM Transactions on Computational Biology and Bioinformatics 8(2) (2011)

[2] Kim, J.K., Choi, S.: Probabilistic models for semi-supervised discriminative motif discovery in DNA sequences. IEEE/ACM Transactions on Computational Biology and Bioinformatics 8(5) (2011)

[3] Vens, C., Rosso, M.N., Danchin, E.G.J.: Identifying discriminative classifcation-based motifs in biological sequences. Bioinformatics 27(9), 1231–1238 (2011)

[4] Gao, M., Nettles, R.E., et al.: Chemical genetics strategy identifies an HCV NS5A inhibitor with a potent clinical effect. Nature 465, 953–960 (2010)

[5] Guilou-Guillemette, H.L., Vallet, S., Gaudy-Graffin, C., Payan, C., Pivert, A., Goudeau, A., Lunel-Fabiani, F.: Genetic diversity of the hepatitis C virus: Impact and issues in the antiviral therapy. World Journal of Gastroenterology 13(17), 2416–2426 (2007)

[6] Los Alamos National Laboratory, http://hcv.lanl.gov/

[7] Genbank, http://www.ncbi.nlm.nih.gov/genbank/
[8] Hepatitis Virus Database, http://s2as02.genes.nig.ac.jp/
[9] Ho, T.B., Kawasaki, S., Le, N.T., Kanda, T., Le, N., Takabayashi, K., Yokosuka, O.: Finding HCV NS5A discriminative motifs for assessment of INF/RBV therapy effect. In: European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases (2011)
[10] Redhead, E., Bailey, T.L.: Discriminative motif discovery in DNA and protein sequences using the DEME algorithm. BMC Bioinformatics 8 (2007)
[11] Sinha, S.: Discriminative motifs. Journal of Computational Biology 10 (2003)
[12] Zhu, X.: Tutorial on semi-supervised learning - ICML (2007)

Comparison of Cost for Zero-One and Stage-Dependent Fuzzy Loss Function

Robert Burduk

Department of Systems and Computer Networks, Wroclaw University of Technology,
Wybrzeze Wyspianskiego 27, 50-370 Wroclaw, Poland
robert.burduk@pwr.wroc.pl

Abstract. In the paper we consider the two-stage binary classifier based on Bayes rule. Assuming that both the tree structure and the feature used at each non-terminal node have been specified, we present the expected total cost. This cost is considered for two types of loss function. First is the zero-one loss function and second is the node-dependent fuzzy loss function. The work focuses on the difference between the expected total costs for these two cases of loss function in the two-stage binary classifier. The obtained results are presented on the numerical example.

1 Introduction

The problem of cost-sensitive classification is broadly discussed in literature. The major costs are the costs of feature measurements (costs of tests) and the costs of classification errors. The costs of feature measurements are discussed in [10], [11], [12], [14] other works included the cost of classification errors [1], [8]. In a more realistic setup, there are good reasons for considering both the costs of feature measurements and the costs of classification errors. For example, there should be a balance between the cost of measuring each feature and the contribution of the test to accurate classification. It often happens that the benefits of further classification are not worth the costs of feature measurements. This means that a cost must be assigned to both the tests and the classification errors. In the works [3], [17], [13], [15] both feature costs and misclassification costs are taken into account.

The class of the fuzzy-valued loss functions is definitely much wider than the class of the real-valued ones [6]. This fact reflects the richness of the fuzzy expected loss approach to describe the consequences of incorrect classification in contrast with the real-valued approach. For this reason, several studies have previously developed decision problems in which values assessed to the consequences of decisions are assumed to be fuzzy [4], [7], [16].

The synthesis of the hierarchical classifier is a complex problem. It involves specifying the following components:

- decision logic, i.e. hierarchical ordering of classes,
- feature used at each stage of decision,
- decision rules (strategy) of performing the classification.

J.-S. Pan, S.-M. Chen, N.T. Nguyen (Eds.): ACIIDS 2012, Part I, LNAI 7196, pp. 385–392, 2012.

In this paper, we consider the costs in two-stage hierarchical classifier. This means that we will deal only with the presentation of costs for different loss function, assuming that both the tree skeleton and the feature used at each non-terminal node have been specified. In our study we consider the zero-one and node-dependent fuzzy loss functions. We compared the costs in two-stage hierarchical classifier for these two cases of loss functions.

The content of the work is as follows: Section 2 introduces the necessary background and describes the hierarchical classifier. In section 3, the recognition algorithms for two cases of loss function are presented. In section 4, we present the cost model and the difference between the costs of the two-stage hierarchical classifier with the zero-one and node-dependent fuzzy loss functions.

2 Decision Problem Statement

Let us consider a pattern recognition problem in which the number of classes equals 4. Let us assume that the classes are organized in a two-stage binary decision tree. Let us number all the nodes of the constructed decision-tree with consecutive numbers of $0, 1, 2, \ldots$, reserving 0 for the root-node and let us assign numbers of classes from $\mathcal{M} = \{1, 2, 3, 4\}$ set to terminal nodes so that each one of them can be labelled with the number of the class connected with that node. This allows us to introduce the following notation [2]:

- $\mathcal{M}(n)$ – the set of numbers of nodes, which distance from the root is n, $n = 0, 1, 2$. In particular $\mathcal{M}(0) = \{0\}$, $\mathcal{M}(N) = \mathcal{M}$,
- $\overline{\mathcal{M}} = \bigcup\limits_{n=0}^{N-1} \mathcal{M}(n)$ – the set of interior node numbers (non terminal),
- $\mathcal{M}_i \subseteq \mathcal{M}(N)$ – the set of class labels attainable from i-th node $(i \in \overline{\mathcal{M}})$,
- $\mathcal{M}^i$ – the set of numbers of immediate descendant nodes $(i \in \overline{\mathcal{M}})$,
- m_i – the number of direct predecessor of i-th node $(i \neq 0)$.

We will continue to adopt the probabilistic model of the recognition problem, i.e. we will assume that the class label of the pattern being recognized $j_N \in \mathcal{M}(N)$ and its observed features x are realizations of a couple of random variables $\boldsymbol{J}_N$ and $\boldsymbol{X}$. The complete probabilistic information denotes the knowledge of a priori probabilities of classes:

$$p(j_N) = P(J_N = j_N), \quad j_N \in \mathcal{M}(N) \tag{1}$$

and class-conditional probability density functions:

$$f_{j_N}(x) = f(x/j_N), \quad x \in X, \quad j_N \in \mathcal{M}(N) \ . \tag{2}$$

Let us

$$x_i \in X_i \subseteq R^{d_i}, \quad d_i \leq d, \quad i \in \mathcal{M} \tag{3}$$

denote the vector of features used at i-th node, which have been selected from vector x.

Our aim is now to calculate the so-called multistage recognition strategy $\pi_N = \{\Psi_i\}_{i \in \overline{\mathcal{M}}}$, that is the set of recognition algorithms in the form:

$$\Psi_i : X_i \to \mathcal{M}^i, \quad i \in \overline{\mathcal{M}} \ . \tag{4}$$

Formula (4) is a decision rule (recognition algorithm) used at i-th node which maps observation subspace to the set of immediate descendant nodes of i-th node. Equivalently, decision rule (4) partitions observation subspace X_i into disjoint decision regions $D_{x_i}^k$, $k \in \mathcal{M}^i$, so that observation x_i is allocated to node k if $k_i \in D_{x_i}^k$, namely:

$$D_{x_i}^k = \{x_i \in X_i : \Psi_i(x_i) = k\}, \ k \in \mathcal{M}^i, \quad i \in \overline{\mathcal{M}}. \tag{5}$$

Let us $L(i_N, j_N)$ denote the loss incurred when the object of class j_N is assigned to class i_N $(i_N, j_N \in \mathcal{M}(N))$. Our aim is to minimize the mean risk, that is the mean value of the loss function:

$$R^*(\pi_N^*) = \min_{\Psi_{i_n}, \Psi_{i_{N-1}}} R(\pi_N) = \min_{\Psi_{i_n}, \Psi_{i_{N-1}}} E[L(I_N, J_N)]. \tag{6}$$

We will refer to π_N^* strategy as the globally optimal N-stage recognition strategy.

3 The Recognition Algorithm

3.1 Zero-One Loss Function

Let us assume the zero-one loss function:

$$L(i_N, j_N) = I(i_N \neq j_N). \tag{7}$$

This loss function assigns loss equal 0 for the correct classification and equal 1 for the incorrect classification.

By putting (7) into (6), we obtain the decision rules for the two-stage decision tree:

$$\begin{array}{c} \Psi_{i_n}^*(x_{i_n}) = i_{n+1}, \\ \displaystyle\sum_{j_N \in \mathcal{M}_{i_{n+1}}} q^*(j_N / i_{n+1}, j_N) p(j_N) f_{j_N}(x_{i_n}) = \\ = \displaystyle\max_{k \in \mathcal{M}^{i_n}} \sum_{j_N \in \mathcal{M}_k} q^*(j_N / k, j_N) p(j_N) f_{j_N}(x_{i_n}) \end{array}, \tag{8}$$

where $q^*(j_N / i_{n+1}, j_N)$ denotes the probability of accurate classification of the object of class j_N at the second stage using π_N^* strategy rules, on condition that at the first stage i_{n+1} decision has been made.

3.2 Node-Dependent Fuzzy Loss Function

Let us assume now:

$$\widetilde{L}(i_N, j_N) = \widetilde{L}_w \tag{9}$$

where w is the first common predecessor of nodes i_N and j_N. The fuzzy loss function defined as above means that the loss depends on the node of the decision tree at which misclassification has been made. The interpretation of this loss function is presented in Fig. 1.

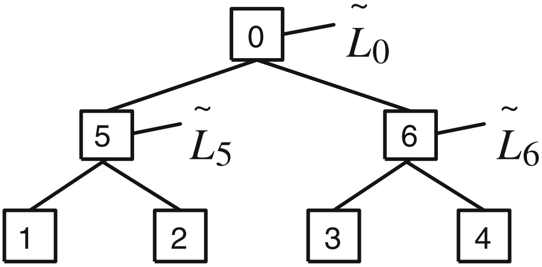


Fig. 1. Interpretation of node-dependent fuzzy loss function

By putting (9) into (6), we obtain the optimal (Bayes) strategy whose decision rules at the first stage are as follows:

$$\begin{array}{c} \Psi^*_{i_n}(x_{i_n}) = i_{n+1}, \\ (\widetilde{L}_0 - \widetilde{L}_{i_{n+1}}) p(i_{n+1}) f_{i_{n+1}}(x_{i_n}) + \\ + \widetilde{L}_{i_{n+1}} \sum\limits_{j_N \in \mathcal{M}^{j_{N-1}}} [q^*(j_N/i_{n+1}, j_N) p(j_N) f_{j_N}(x_{i_n})] = \\ = \max\limits_{k \in \mathcal{M}^{i_n}} \Big\{ (\widetilde{L}^S_0 - \widetilde{L}_k) p(k) f_k(x_{i_n}) + \\ + \widetilde{L}_k \sum\limits_{j_N \in \mathcal{M}^k} [q^*(j_N/k, j_N) p(j_N) f_{j_N}(x_{i_n})] \Big\} \end{array} , \tag{10}$$

where $q^*(j_N/i_{n+1}, j_N)$ denotes the probability of accurate classification of the object of class j_N at the second stage using π^*_N strategy rules, on condition that at the first stage i_{n+1} decision has been made.

Decision rules at the second stage of classification are like for the single-stage classifier with zero-one loss function.

4 Cost Model

In the costs model we specify two costs. There are the feature acquisition costs and the misclassification costs. We assume that the feature acquisition cost for each internal node is known. It means that i-th node has associated feature acquisition costs $FAC(i)$. In this case, each feature has an independent cost and the cost of a set of features is just an additive cost.

Each patch $s(k), k \in \mathcal{M}$ represents a sequence of ordered feature values and the final classification. Each path has an associated feature acquisition cost. It can be computed as follows:

$$FAC(s(k)) = \sum_{i \in s(k), \quad i \notin \mathcal{M}} FAC(i). \tag{11}$$

Now we present the misclassification cost. Each path has an associated expected misclassification cost. It can be computed as follows:

$$EMC(s(k)) = \sum_{i_N \in \mathcal{M}(N)} L(i_N, k) \prod_{i_n \in s(k) - \{0\}} q(i_n/m_{i_n}, k). \tag{12}$$

The total costs of a path is the sum of the feature acquisition costs and the expected misclassification costs:

$$TC(s(k)) = FAC(s(k)) + EMC(s(k)). \tag{13}$$

The expected total cost of the globally optimal strategy π_N^* is then the sum of the total cost of each path, weighted by the probability of the following path:

$$ETC(\pi_N^*) = \sum_{k=1}^{\mathcal{M}} P(s(k))TC(s(k)). \tag{14}$$

4.1 Cost for Two-Stage Binary Classifier

For two-stage binary classifier the expected misclassification cost can be presented as follows:

$$EMC(s(k)) = L(m_k)q(m_k/r, k)q(l/m_k, k) + L(r)q(n/r, k), \tag{15}$$

where $L(m_k)$ is the loss function for the immediate predecessor of k-th node, $L(r)$ is the loss function for the root node, $q(m_k/r, k))$ is the probability of correct classification in the root node r on condition that the correct class is k, $q(l/m_k, k)$ is the probability of error in m_k node on condition that the correct class is k, $q(n/m_k, k)$ is the probability of error in root node on condition that the correct class is k. Additionally $l \in \mathcal{M}^{m_k}, l \neq k$ and $n \in \mathcal{M}^r, n \neq m_k$.

For the zero-one loss function the values of $L(m_k)$ and $L(r)$ are equal 1. For the node-dependent fuzzy loss function $L(r) = \widetilde{L}(r)$ is the loss function in the root node and $L(m_k) = \widetilde{L}(m_k)$ is the loss function on the node of the decision tree in the second stage.

For the zero-one loss function the expected misclassification cost can be presented as follows:

$$\begin{aligned} EMC_{01}(s(k)) &= q(m_k/r, k)(1 - q(k/m_k, k)) + q(n/r, k) = \\ &= q(m_k/r, k) + q(n/r, k) - q(m_k/r, k)q(k/m_k, k) = \\ &= 1 - q(m_k/r, k)q(k/m_k, k). \end{aligned} \tag{16}$$

For the node-dependent loss function we can assume that the loss function in the root node (first stage of the decision tree) is larger than at the loss in each node of the second stage $\widetilde{L}(r) = \widetilde{L}(r) + a(m_k)$, where $a(m_k)$ presents the additional value of loss in node m_k. Such an assumption is quite natural, means that we make the greater error in classification sooner. For this assumption the expected misclassification cost for the node-dependent fuzzy loss function can be presented as follows:

$$\begin{aligned} EMC_{ndf}(s(k)) &= \widetilde{L}(m_k)q(m_k/r,k)(1-q(k/m_k,k)) + \widetilde{L}(r)q(n/r,k) = \\ &= \widetilde{L}(m_k)q(m_k/r,k) + (\widetilde{L}(m_k) + a(m_k))q(n/r,k) - \\ &-\widetilde{L}(m_k)q(m_k/r,k)q(k/m_k,k) = \\ &= \widetilde{L}(m_k) + a(m_k)q(n/r,k) - \\ &-\widetilde{L}(m_k)q(m_k/r,k)q(k/m_k,k). \end{aligned} \tag{17}$$

Now we will present the difference between the expected total cost for the zero-one and the node-dependent loss function. In these two cases the feature acquisition costs are the same. Using equations (16) and (17) the difference in the expected total cost can be represented as:

$$\begin{aligned} ETC_{ndf}(\pi_N^*) & -ETC_{01}(\pi_N^*) = \\ & = \sum_{k=1}^{\mathcal{M}} P(s(k))(EMC_{ndf}(s(k)) - EMC_{01}) = \\ & = \sum_{k=1}^{\mathcal{M}} P(s(k))(\widetilde{L}(m_k) - 1 + a(m_k)q(n/r,k) + \\ & +(1-\widetilde{L}(m_k))q(m_k/r,k)q(k/m_k,k)). \end{aligned} \tag{18}$$

The obtained result have not simple interpretation. The final result depends on the defuzzification and comparison of the fuzzy numbers method because in order to obtain the final result we must perform operations on fuzzy numbers and then get the result in the domain of real numbers. The obtained results will be presented in the form of an example.

4.2 Illustrative Example

Let us consider the two-stage binary classifier in Fig. 1. Four classes have identical a priori probabilities which equal 0.25. Class-conditional probability density functions of features $\mathbf{X}_0$, $\mathbf{X}_5$ and $\mathbf{X}_6$ are normally distributed in each class with the following class-conditional probability density functions: $f_1(x_0) = f_2(x_0) = f_5(x_0) = N(2,1)$, $f_3(x_0) = f_4(x_0) = f_6(x_0) = N(4,1)$, $f_1(x_5) = N(3,2)$, $f_2(x_5) = N(7,2)$, $f_3(x_6) = N(0,2)$, $f_4(x_6) = N(4,2)$. The node-dependent fuzzy loss functions are as follows: $\widetilde{L}_0 = \widetilde{L}(m_k) = (1.5, 2, 2.5)_T$, $\widetilde{L}_5 = (1.5, 1.5, 2)_T$ and $\widetilde{L}_6 = (1, 1, 1.5)_T$. These functions are described by the triangular fuzzy numbers.

The decision regions for the case of the zero-one and the node-dependent fuzzy loss function are the same and are respectively: $D_{x_5}^{*(1)} \subset (-\infty, 5)$, $D_{x_5}^{*(2)} \subset (5, \infty)$, $D_{x_6}^{*(3)} \subset (-\infty, 2)$ and $D_{x_6}^{*(4)} \subset (2, \infty)$. From (8) we obtain the decision region at the first stage for the zero-one loss function: $D_{x_0}^{*(5)} \subset (-\infty, 3)$, $D_{x_0}^{*(6)} \subset (3, \infty)$. For the node-dependent fuzzy loss function we use the subjective λ-method for comparing fuzzy risks [5]. Now from (10) we obtain the decision region at the first stage $D_{x_0}^{*(5)} \subset (-\infty, 2.97)$, $D_{x_0}^{*(6)} \subset (2.97, \infty)$ and $D_{x_0}^{*(5)} \subset (-\infty, 2.98)$, $D_{x_0}^{*(6)} \subset$

$(2.98, \infty)$ for $\lambda = 0$ and $\lambda = 1$ respectively. For the present example the difference between the expected total cost for the zero-one and the node-dependent loss function (18) is equal 0.23 and 0.22 for $\lambda = 0$ and $\lambda = 1$ respectively. For these calculations deffuzification center of gravity method was used [9].

5 Conclusion

In this paper, we have concentrated on the costs of the two stage binary hierarchical classifier. In this study we assume that the decision tree is known, that is, the work does not generate its structure. The study considered two types of costs, the feature acquisition costs and the misclassification costs. For the zero-one and the node-dependent fuzzy loss function an expected total cost of the globally optimal strategy was presented. The work focuses on the difference between the expected total cost for these two cases of the loss function.

In future work we can consider various defuzzification and comparison of the fuzzy numbers method in order to obtain the final result of the discussed two types loss functions.

Acknowledgements. This work is supported in part by The Polish Ministry of Science and Higher Education under the grant which is being realized in years 2010-2013.

References

1. Breiman, L., Friedman, J., Olshen, R., Stone, C.: Classification and regression trees. California, Wadsworth (1984)
2. Burduk, R.: Classification error in Bayes multistage recognition task with fuzzy observations. Pattern Analysis and Applications 13(1), 85–91 (2010)
3. Burduk, R.: Costs-Sensitive Classification in Multistage Classifier with Fuzzy Observations of Object Features. In: Corchado, E., Kurzyński, M., Woźniak, M. (eds.) HAIS 2011, Part II. LNCS (LNAI), vol. 6679, pp. 245–252. Springer, Heidelberg (2011)
4. Burduk, R., Kurzynski, M.: Two-stage binary classifier with fuzzy-valued loss function. Pattern Analysis and Applications 9(4), 353–358 (2006)
5. Campos, L., González, A.: A Subjective Approach for Ranking Fuzzy Numbers. Fuzzy Sets and Systems 29, 145–153 (1989)
6. Domingos, P.: MetaCost: A General Method for Making Classifiers Cost-Sensitive. In: Proceedings of the Fifth International Conference on Knowledge Discovery and Data Mining (KDD 1999), pp. 155–164 (1999)
7. Jain, R.: Decision-Making in the Presence of Fuzzy Variables. IEEE Trans. Systems Man and Cybernetics 6, 698–703 (1976)
8. Knoll, U., Nakhaeizadeh, G., Tausend, B.: Cost-Sensitive Pruning of Decision Trees. In: Bergadano, F., De Raedt, L. (eds.) ECML 1994. LNCS, vol. 784, pp. 383–386. Springer, Heidelberg (1994)
9. Van Leekwijck, W., Kerre, E.: Defuzzification: criteria and classification. Fuzzy Sets and Systems 108(2), 159–178 (1999)

10. Núñez, M.: The use of background knowledge in decision tree induction. Machine Learning 6(3), 231–250 (1991)
11. Penar, W., Woźniak, M.: Experiments on classifiers obtained via decision tree induction methods with different attribute acquisition cost limit. Advances in Soft Computing 45, 371–377 (2007)
12. Penar, W., Woźniak, M.: Cost-sensitive methods of constructing hierarchical classifiers. Expert Systems 27(3), 146–155 (2010)
13. Saar-Tsechansky, M., Melville, P., Provost, F.: Active feature-value acquisition. Management Science 55(4), 664–684 (2009)
14. Tan, M.: Cost-sensitive learning of classification knowledge and its applications in robotics. Machine Learning 13, 7–33 (1993)
15. Turney, P.: Cost-sensitive classification: Empirical evaluation of a hybrid genetic decision tree induction algorithm. Journal of Artificial Intelligence Research 2, 369–409 (1995)
16. Viertl, R.: Statistical Methods for Non-Precise Data. CRC Press, Boca Raton (1996)
17. Yang, Q., Ling, C., Chai, X., Pan, R.: Test-cost sensitive classification on data with missing values. IEEE Transactions on Knowledge and Data Engineering 18(5), 626–638 (2006)

Investigation of Rotation Forest Ensemble Method Using Genetic Fuzzy Systems for a Regression Problem

Tadeusz Lasota[1], Zbigniew Telec[2], Bogdan Trawiński[2], and Grzegorz Trawiński[3]

[1] Wrocław University of Environmental and Life Sciences, Dept. of Spatial Management
ul. Norwida 25/27, 50-375 Wrocław, Poland
[2] Wrocław University of Technology, Institute of Informatics,
Wybrzeże Wyspiańskiego 27, 50-370 Wrocław, Poland
[3] Wrocław University of Technology, Faculty of Electronics,
Wybrzeże S. Wyspiańskiego 27, 50-370 Wrocław, Poland
tadeusz.lasota@up.wroc.pl, grzegorztrawinski@wp.pl,
{zbigniew.telec,bogdan.trawinski}@pwr.wroc.pl

Abstract. The rotation forest ensemble method using a genetic fuzzy rule-based system as a base learning algorithm was developed in Matlab environment. The method was applied to the real-world regression problem of predicting the prices of residential premises based on historical data of sales/purchase transactions. The computationally intensive experiments were conducted aimed to compare the accuracy of ensembles generated by our proposed method with bagging, repeated holdout, and repeated cross-validation models. The statistical analysis of results was made employing nonparametric Friedman and Wilcoxon statistical tests.

Keywords: rotation forest, genetic fuzzy systems, bagging, repeated holdout, cross-validation, ensemble models, property valuation.

1 Introduction

Ensemble machine learning models have been focusing attention of many researchers due to its ability to reduce bias and/or variance when compared to their single model counterparts. The ensemble learning methods combine the output of machine learning systems, called in literature “weak learners” due to their performance, in order to get smaller prediction errors (in regression) or lower error rates (in classification). The individual estimator must provide different patterns of generalization, thus the diversity plays a crucial role in the training process. To the most popular methods belong bagging [1], boosting [19], and stacking [20]. In this paper we focus on bagging family of techniques.

Bagging, which stands for bootstrap aggregating, devised by Breiman [1] is one of the most intuitive and simplest ensemble algorithms providing good performance. Diversity of learners is obtained by using bootstrapped replicas of the training data. That is, different training data subsets are randomly drawn with replacement from the original training set. So obtained training data subsets, called also bags, are used then

J.-S. Pan, S.-M. Chen, N.T. Nguyen (Eds.): ACIIDS 2012, Part I, LNAI 7196, pp. 393–402, 2012.

to train different classification and regression models. Theoretical analyses and experimental results proved benefits of bagging, especially in terms of stability improvement and variance reduction of learners for both classification and regression problems [4], [7].

Another approach to ensemble learning is called the random subspaces, also known as attribute bagging [3]. This approach seeks learners diversity in feature space subsampling. All component models are built with the same training data, but each takes into account randomly chosen subset of features bringing diversity to ensemble. For the most part, feature count is fixed at the same level for all committee components. The method is aimed to increase generalization accuracies of decision tree-based classifiers without loss of accuracy on training data. Ho showed that random subspaces can outperform bagging and in some cases even boosting [9]. While other methods are affected by the curse of dimensionality, random subspace technique can actually benefit out of it.

Both bagging and random subspaces were devised to increase classifier or regressor accuracy, but each of them treats the problem from different point of view. Bagging provides diversity by operating on training set instances, whereas random subspaces try to find diversity in feature space subsampling. Breiman [2] developed a method called random forest which merges these two approaches. Random forest uses bootstrap selection for supplying individual learner with training data and limits feature space by random selection. Some recent studies have been focused on hybrid approaches combining random forests with other learning algorithms [8], [12].

Rodríguez et al. [18] proposed in 2006 a new classifier ensemble method, called rotation forest, applying Principal Component Analysis (PCA) to rotate the original feature axes in order to obtain different training sets for learning base classifiers. Their approach attempted to achieve simultaneously greater diversity and accuracy of components within the ensemble. The diversity was obtained by applying PCA to some kind of feature manipulation for individual classifier and accuracy was acquired by holding all principal components (i.e. all features) and employing all instances to train each classifier. They conducted a validation experiment comparing rotation forest with bagging, AdaBoost, and random forests. The authors explore the effect of the design choices and parameter values on the performance of rotation forest ensembles in [15]. Zhang & Zhang successfully combined rotation forest with other ensemble classification techniques, namely with bagging [21] and with AdaBoost [22] and tested them using benchmark classification datasets. Kotsianis [12] built a hybrid ensemble combing bagging, boosting, rotation forest and random subspace. Jędrzejowicz & Jędrzejowicz [10] explored rotation forest based classifier ensembles created using expression trees induced by gene expression programming. Some research was also done into the application of rotation forest into regression problems. Zhang et al. [23] compared the method with bagging, random forest, AdaBoost.R2, and a single regression tree. Kotsiantis & Pintelas [13] proposed a combined technique of local rotation forest of decision stumps.

In the majority of the aforementioned experiments on rotation forest techniques various kinds of decision trees were used as base learning algorithms because they reveal relative high computational efficiency and provide diverse component models of an ensemble. Due to the fact that in rotation forest each tree is trained on a dataset in a rotated feature space and it builds classification regions using hyperplanes parallel to the feature axes, even a small rotation of the axes may lead to a very different tree [15]. In this paper we explore the rotation forest technique using a genetic fuzzy system as a base learning method applied to the real-world regression problem of predicting the prices of residential premises based on historical data of sales/purchase transactions obtained from a cadastral system. To the best of our knowledge there are not yet any published results of ensemble models created using rotation forest with genetic fuzzy systems, probably due to their considerable computational load. The research was conducted with our newly developed experimental system in Matlab environment to test multiple models using different resampling methods. So far, we have investigated genetic fuzzy systems and genetic fuzzy networks applied to construct regression ensemble models to assist with real estate appraisal using our system [11], [16], [17].

2 Rotation Forest with Genetic Fuzzy System

Our study consisted in the application of the rotation forest (RF) method using a genetic fuzzy system (GFS) as a base learning algorithm to a real-world regression problem of predicting the prices of residential premises based on historical data of sales/purchase transactions obtained from a cadastral system. We compared in terms of accuracy the models produced by RF with the ones generated employing other resampling techniques such repeated bootstrap, i.e. bagging (BA), repeated holdout (HO), and repeated cross-validation (CV).

In our Matlab experimental system we developed a data driven fuzzy system of Mamdani type, where for each input and output variables triangular and trapezoidal membership functions were automatically determined by the symmetric division of the individual attribute domains. The evolutionary optimization process utilized a genetic algorithm of Pittsburgh type and combined both learning the rule base and tuning the membership functions using real-coded chromosomes. Similar designs are described in [5], [6], [14]. The rotation forest method was implemented in the Matlab environment based on ideas described by Rodrígues et al. [18] and Zhang et al. [23]. The pseudo code of our rotation forest ensemble method employing the genetic fuzzy system as a base learning algorithm is presented in Fig. 1.

3 Plan of Experiments

Four following approaches to create ensemble models using genetic fuzzy systems as a base learning algorithm were employed in our experiments: rotation forest (RF), bagging (BA), repeated holdout (HO), and repeated cross-validation (CV). In each case an ensemble comprised 50 component models. They were tested with following parameters:

RF: M=2, 3, and 4 – rotation forest with three different numbers of input attributes in individual feature subset, as described in Section 2.

BA: B90, B70, B50 – rotation forest with three different sizes of bootstrap samples drawn from X_{ij} equal to 50%, 70%, and 90%, respectively, as described in Section 2.

```
Given:
  • GFS: genetic fuzzy system used as a base learning algorithm
  • L: number of fuzzy models (fuzzy inference systems - FIS) that make up an ensemble
  • n: number of input attributes in a base training data set
  • N: number of instances in a base training data set
  • X: N × n matrix of input attribute values for individual instances
  • Y: N × 1 vector of output attribute values
  • $\mathcal{T}$=(X, Y): base training dataset as the concatenation of X and Y
  • F: feature set, $F_{ij}$ – j-th attribute subset used for training i-th $FIS_i$
  • M: number of input attributes in individual feature subset
  • $X_{ij}$: data corresponding $F_{ij}$ selected from matrix $X$
  • $X'_{ij}$ : - bootstrap sample from $X_{ij}$
  • $D_{ij}$: M× M matrix to store the coefficients of principal components computed by PCA
  • $R_i$: block diagonal matrix built of $D_{ij}$ matrices
  • $R_i^a$ : rotation matrix to obtain training set for i-th FIS

Training Phase
For i=1,2,   ,L
  • Calculate rotation matrix $R_i^a$ for i-th FIS
      1. Randomly split F into K subsets $F_{ij}$ (j=1,..K) for M attributes each (last subset may
         contain less than M attributes)
      2. For j=1,2,…,K
           a. Select columns of X that correspond to the attributes of $F_{ij}$ to compose a
              new matrix $X_{ij}$
           b. Draw a bootstrap sample $X'_{ij}$ from $X_{ij}$, with sample size smaller than
              that of $X_{ij}$, with eg size (B) equal to 50%, 70%, or 90% of $X_{ij}$
           c. Apply PCA to $X'_{ij}$ to obtain a matrix $D_{ij}$ whose p-th column comprises
              the coefficients of p-th principal component
      3. EndFor
      4. Build a block diagonal matrix $R_i$ of matrices $D_{ij}$ (j=1,2,…,K)
      5. Construct the resulting rotation matrix $R_i^a$ by rearranging rows of $R_i$ to match the
         order of attributes in F
  • Compute (X $R_i^a$ , Y) and use it as an input of GFS (training dataset) to obtain i-th $FIS_i$
EndFor

Predicting Phase
  • For any instance $\mathbf{x}_t$ from test dataset, let $FIS_i(x_t R_i^a)$ be the value predicted by i-th
    fuzzy model, then the predicted output value $y_t^p$ can be computed as

              $y_t^p = \frac{1}{L}\sum_{i=1}^{L} FIS_i(x_t R_i^a)$
```

Fig. 1. Pseudo code of rotation forest ensemble method employing the genetic fuzzy system

BA: B100, B80 – m-out-of-n bagging with replacement with different sizes of samples. The numbers in the codes indicate what percentage of the training set was drawn.

HO: H100, H80 – repeated holdout, i.e., m-out-of-n bagging without replacement with different sizes of samples. The numbers in the codes indicate what percentage of the training set was drawn.

CV: 5x10cv, 10x5cv– repeated cross-validation, with different k-fold cross-validation splits, for *k=5 and 10*, were repeated 5 and 10 times, respectively.

Real-world data used in experiments was drawn from an unrefined dataset containing above 50 000 records referring to residential premises transactions accomplished in one Polish big city with the population of 640 000 within eleven years from 1998 to 2008. In this period most transactions were made with non-market prices when the council was selling flats to their current tenants on preferential terms. First of all, transactional records referring to residential premises sold at market prices were selected. Then, the dataset was confined to sales transaction data of apartments built before 1997 and where the land was leased on terms of perpetual usufruct. Hence, the final dataset counted 5303 records. Five following attributes were pointed out as price drivers by professional appraisers: usable area of a flat (*Area*), age of a building construction (*Age*), number of storeys in the building (Storeys), number of rooms in the flat including a kitchen (*Rooms*), the distance of the building from the city centre (*Centre*), in turn, price of premises (*Price*) was the output variable.

Due to the fact that the prices of premises change substantially in the course of time, the whole 11-year dataset cannot be used to create data-driven models. In order to obtain comparable prices it was split into subsets covering individual years. Then the prices of premises were updated according to the trends of the value changes over 11 years. Starting from the beginning of 1998 the prices were updated for the last day of subsequent years. The trends were modelled by polynomials of degree three. The chart illustrating the change trend of average transactional prices per square metre is given in Fig. 2. We might assume that one-year datasets differed from each-other and might constitute different observation points to compare the accuracy of ensemble models in our study. The sizes of one-year datasets are given in Table 1.

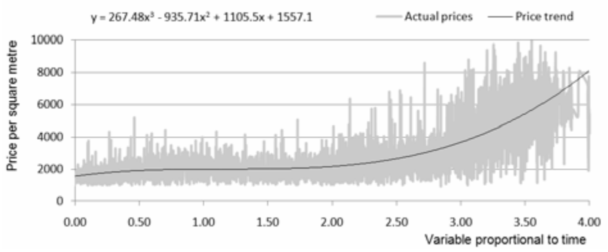


Fig. 2. Change trend of average transactional prices per square metre over time

Table 1. Number of instances in one-year datasets

1998	1999	2000	2001	2002	2003	2004	2005	2006	2007	2008
269	477	329	463	530	653	546	580	677	575	204

As a performance function the root mean square error (RMSE) was used, and as aggregation functions of ensembles arithmetic averages were employed. Each input and output attribute in individual dataset was normalized using the min-max approach. The parameters of the architecture of fuzzy systems as well as genetic algorithms are listed in Table 2.

Table 2. Parameters of GFS used in experiments

Fuzzy system	Genetic Algorithm
Type of fuzzy system: Mamdani	Chromosome: rule base and mf, real-coded
No. of input variables: 5	Population size: 50
Type of membership functions (mf): triangular	Fitness function: MSE
No. of input mf: 3	Selection function: tournament
No. of output mf: 5	Tournament size: 4
No. of rules: 15	Elite count: 2
AND operator: prod	Crossover fraction: 0.8
Implication operator: prod	Crossover function: two point
Aggregation operator: probor	Mutation function: custom
Defuzzyfication method: centroid	No. of generations: 100

In order to ensure the same experimental conditions for all ensemble learning methods, before employing a given method each one-year dataset was split into training and test sets in the proportion of 80% to 20% instances using stratified approach. Namely, each dataset was divided into several clusters with k-means algorithm. The number of clusters was determined experimentally according to the best value of Dunn's index. Then, 80% of instances from each cluster were randomly selected to the training set and remaining 20% went to the test set. After that, each learning method was applied to the training set and 50 models were generated. The performance of each model was determined using test set. And finally, the average value of accuracy measure over all component models constituted the resulting indicator of the performance of a given method over the one-year dataset. A schema illustrating the flow of experiments is shown in Fig. 3. In turn, having averaged values of RMSE for all one-year datasets we were able to make non-parametric Friedman and Wilcoxon statistical tests to compare the performance of individual methods with different parameters as well as all considered methods.

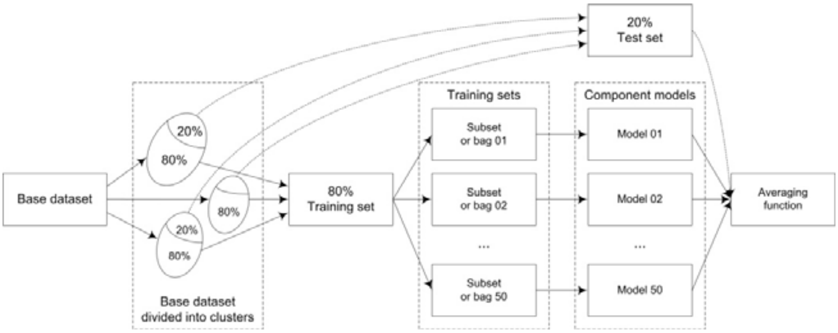


Fig. 3. Outline of experiment with training and test sets created using stratified approach

4 Results of Experiments

The performance of RF for different values of parameters M and B in terms of RMSE is illustrated graphically in Figures 4 and 5 respectively. In turn, the results provided by the BA, HO, and CV models created using genetic fuzzy systems (GFS) are shown in Figures 6, 7, and 8 respectively. The comparison of the best selected variants of individual methods is given in Fig. 9. The statistical analysis of the results was carried out with non-parametric Friedman and Wilcoxon tests performed in respect of RMSE values of all models built over 11 one-year datasets. The tests were made using Statisitca package. Average ranks of individual models provided by Friedman tests are shown in Table 3, where the lower rank value the better model, and asterisks indicate statistical significance at the level α=0.05.

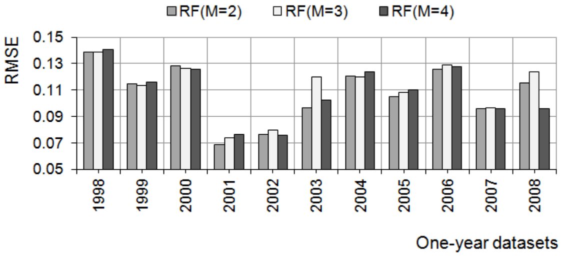


Fig. 4. Performance of *Rotation Forest* ensembles with different values of parameter M and B70

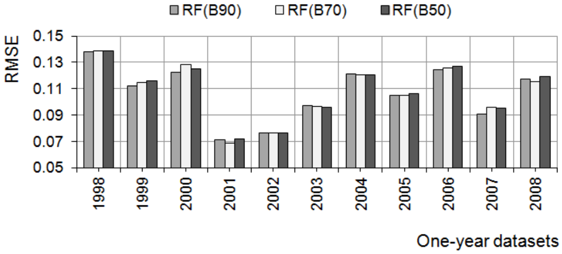


Fig. 5. Performance of *Rotation Forest* ensembles with different values of B and M=2

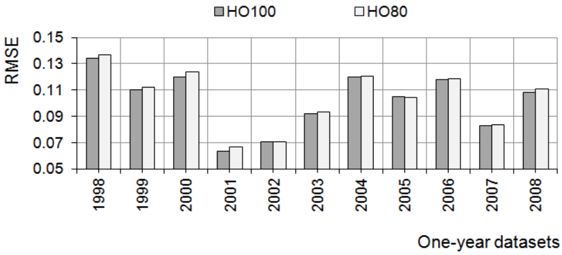


Fig. 6. Performance comparison of *Repeated Holdout* ensembles for H100 and H80

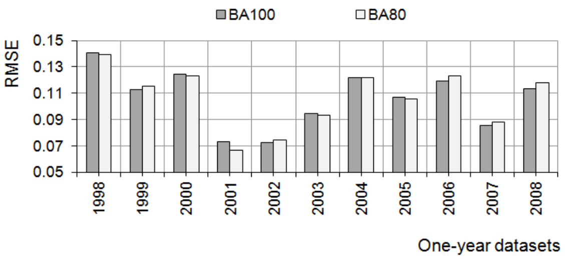


Fig. 7. Performance comparison of *Bagging* ensembles for BA100 and BA80

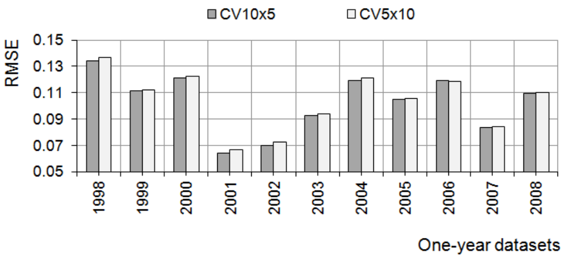


Fig. 8. Performance comparison of *Cross-validation* ensembles for CV10x5 and CV5x10

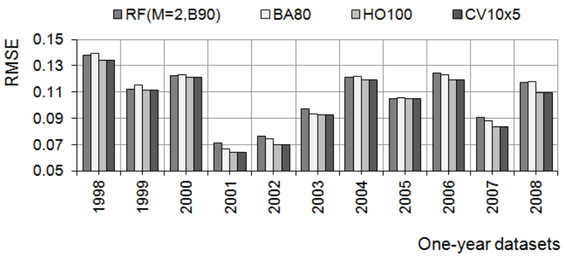


Fig. 9. Performance comparison of the best of *RF, BA, HO, and CV* ensembles

Table 3. Average rank positions of models determined during Friedman test (* significant)

Methods	1st	2nd	3rd	4th
RF for M={2,3,4}	RF(M=2) (1.55)	RF(M=4) (2.18)	RF(M=3) (2.27)	
RF for B={50, 70, 90}	RF(B90) (1.64)	RF(B70) (2.09)	RF(B50) (2.27)	
*HO100, HO80	H100 (1.18)	H80 (1.82)		
BA100, BA80	BA80 (1.45)	BA100 (1.55)		
*CV10x5, 5x10	CV10x5 (1.09)	CV5x10 (1.91)		
*RF,BA,HO, CV	CV10x5 (1.09)	RF(B90) (2.82)	BA80 (3.00)	HO100 (3.09)

With our specific datasets and experimental setup the best results produced RF with M=2 and B90, however the differences in accurracy among RF with different parameters were statictically insignificant. To the final comparison ensembles created using RF with M=2 and B90 were chosen bacause parameter M=2 provides the highest diversity of component models and B90 ensured the best accuracy when compared to HO and BA. As for individual methods, HO100 and CV10x5 revealed significantly better performance than H80 and CV5x10 respectively. No significant differences were observed between B100 and B80.

The final comparison comprised the best selected ensembles produced by individual methods, namely RF(M=2,B90), BA80, H100, and CV10x5. Friedman test indicated significant differences in performance among the methods. According to Wilcoxon test CV outperformed significantly all other techniques. RF, BA, and HO performed equivalently, but among them RF gained the best rank value. When considered each one-year data point separately, RF revealed better performance than BA and HO over 5 one-year datasets and for the dataset of 2005 outperformed all other methods.

5 Conclusions and Future Work

The rotation forest ensemble method using a genetic fuzzy rule-based system system as a base learning algorithm was developed in Matlab environment. The method was applied to the real-world regression problem of predicting the prices of residential premises based on historical data of sales/purchase transactions. The computationally intensive experiments were conducted aimed to compare the accuracy of ensembles generated by our proposed method with bagging, repeated holdout, and repeated cross-validation models. The statistical analysis of results was made employing nonparametric Friedman and Wilcoxon statistical tests.

The overall results of our investigation are as follows. The ensembles created using genetic fuzzy systems revealed prediction accuracy not worse than bagging and repeated holdout models. Cross-validation approach outperformed other techniques. The proposed method seem to be useful for our real-world regression problem. Further investigations into rotation forest combined with genetic fuzzy systems are planned using benchmark regression datasets preprocessed with instance and feature selection algorithms. The resistance of the method to noised data will be also examined.

Acknowledgments. This paper was partially supported by the Polish National Science Centre under grant no. N N516 483840.

References

1. Breiman, L.: Bagging Predictors. Machine Learning 24(2), 123–140 (1996)
2. Breiman, L.: Random Forests. Machine Learning 45(1), 5–32 (2001)
3. Bryll, R.: Attribute bagging: improving accuracy of classifier ensembles by using random feature subsets. Pattern Recognition 20(6), 1291–1302 (2003)
4. Bühlmann, P., Yu, B.: Analyzing bagging. Annals of Statistics 30, 927–961 (2002)
5. Cordón, O., Gomide, F., Herrera, F., Hoffmann, F., Magdalena, L.: Ten years of genetic fuzzy systems: current framework and new trends. Fuzzy Sets and Systems 141, 5–31 (2004)

6. Cordón, O., Herrera, F.: A Two-Stage Evolutionary Process for Designing TSK Fuzzy Rule-Based Systems. IEEE Trans. Sys., Man, and Cyb.-Part B 29(6), 703–715 (1999)
7. Fumera, G., Roli, F., Serrau, A.: A theoretical analysis of bagging as a linear combination of classifiers. IEEE Transactions on Pattern Analysis and Machine Intelligence 30(7), 1293–1299 (2008)
8. Gashler, M., Giraud-Carrier, C., Martinez, T.: Decision Tree Ensemble: Small Heterogeneous Is Better Than Large Homogeneous. In: Seventh International Conference on Machine Learning and Applications, ICMLA 2008, pp. 900–905 (2008)
9. Ho, T.K.: The Random Subspace Method for Constructing Decision Forests. IEEE Transactions on Pattern Analysis and Machine Intelligence 20(8), 832–844 (1998)
10. Jędrzejowicz, J., Jędrzejowicz, P.: Rotation Forest with GEP-Induced Expression Trees. In: O'Shea, J., Nguyen, N.T., Crockett, K., Howlett, R.J., Jain, L.C. (eds.) KES-AMSTA 2011. LNCS (LNAI), vol. 6682, pp. 495–503. Springer, Heidelberg (2011)
11. Kempa, O., Lasota, T., Telec, Z., Trawiński, B.: Investigation of Bagging Ensembles of Genetic Neural Networks and Fuzzy Systems for Real Estate Appraisal. In: Nguyen, N.T., Kim, C.-G., Janiak, A. (eds.) ACIIDS 2011, Part II. LNAI (LNCS), vol. 6592, pp. 323–332. Springer, Heidelberg (2011)
12. Kotsiantis, S.: Combining bagging, boosting, rotation forest and random subspace methods. Artificial Intelligence Review 35(3), 223–240 (2011)
13. Kotsiantis, S.B., Pintelas, P.E.: Local Rotation Forest of Decision Stumps for Regression Problems. In: 2nd IEEE International Conference on Computer Science and Information Technology, ICCSIT 2009, pp. 170–174 (2009)
14. Król, D., Lasota, T., Trawiński, B., Trawiński, K.: Investigation of Evolutionary Optimization Methods of TSK Fuzzy Model for Real Estate Appraisal. International Journal of Hybrid Intelligent Systems 5(3), 111–128 (2008)
15. Kuncheva, L.I., Rodríguez, J.J.: An Experimental Study on Rotation Forest Ensembles. In: Haindl, M., Kittler, J., Roli, F. (eds.) MCS 2007. LNCS, vol. 4472, pp. 459–468. Springer, Heidelberg (2007)
16. Lasota, T., Telec, Z., Trawiński, G., Trawiński, B.: Empirical Comparison of Resampling Methods Using Genetic Fuzzy Systems for a Regression Problem. In: Yin, H., Wang, W., Rayward-Smith, V. (eds.) IDEAL 2011. LNCS, vol. 6936, pp. 17–24. Springer, Heidelberg (2011)
17. Lasota, T., Telec, Z., Trawiński, G., Trawiński, B.: Empirical Comparison of Resampling Methods Using Genetic Neural Networks for a Regression Problem. In: Corchado, E., Kurzyński, M., Woźniak, M. (eds.) HAIS 2011, Part II. LNCS, vol. 6679, pp. 213–220. Springer, Heidelberg (2011)
18. Rodrígeuz, J.J., Kuncheva, I., Alonso, C.J.: Rotation forest: A new classifier ensemble method. IEEE Trans. on Pattern Analysis and Mach. Intel. 28(10), 1619–1630 (2006)
19. Schapire, R.E.: The strength of weak learnability. Mach. Learning 5(2), 197–227 (1990)
20. Wolpert, D.H.: Stacked Generalization. Neural Networks 5(2), 241–259 (1992)
21. Zhang, C.-X., Zhang, J.-S.: A variant of Rotation Forest for constructing ensemble classifiers. Pattern Analysis & Applications 13(1), 59–77 (2010)
22. Zhang, C.-X., Zhang, J.-S.: RotBoost: A technique for combining Rotation Forest and AdaBoost. Pattern Recognition Letters 29(10), 1524–1536 (2008)
23. Zhang, C.-X., Zhang, J.-S., Wang, G.-W.: An empirical study of using Rotation Forest to improve regressors. Applied Mathematics and Computation 195(2), 618–629 (2008)

Data with Shifting Concept Classification Using Simulated Recurrence

Piotr Sobolewski and Michał Woźniak

Department of Systems and Computer Networks, Faculty of Electronics,
Wroclaw University of Technology, Wybrzeże Wyspiańskiego 27,
50-370 Wroclaw, Poland
{piotr.sobolewski,michal.wozniak}@pwr.wroc.pl

Abstract. One of the serious problems of modern pattern recognition is concept drift i.e., model changing during exploitation of a given classifier. The paper proposes how to adapt a single classifier system to the new model without the knowledge of correct classes. The proposed simulated concept recurrence is implemented in the non-recurring concept shift scenario without the drift detection mechanism. We assume that the model could change slightly, what allows us to predict a set of possible models. Quality of the proposed algorithm was estimated on the basis of computer experiment which was carried out on the benchmark dataset.

Keywords: Concept shift, classifier ensemble, simulated recurrence, machine learning.

1 Introduction

Concept shift is a hidden change in the data model, which may be caused e.g., by lack of knowledge about characteristics of data distributions, which may influence the classification accuracy [14]. Let's consider the classification task to name given 3-D objects on the basis of their 2-D images captured from various perspectives. In such scenario the typical classification system might make wrong decisions e.g., if a 3-D cylinder is photographed from different angles, then it can produce different 2-D images, e.g. once being represented by a circle and once by a rectangle. These 2-D figures may also indicate other 3-D objects (a circle may be an image of a sphere or a cone and a rectangle - of a cube or a cuboid). This problem is called "concept drift" and is often categorized as one of the two types: gradual and sudden (also called: concept drift and concept shift, respectively).

1.1 Gradual Concept Drift

Gradual concept drift appears in the data stream as a sequence of minor changes in the hidden features of data, which follow a certain trend. E.g., the data stream is a set of the 2-D images of a certain 3-D object, which starts to rotate. The classification system decides whether the following 2-D images are representing a new 3-D object

J.-S. Pan, S.-M. Chen, N.T. Nguyen (Eds.): ACIIDS 2012, Part I, LNAI 7196, pp. 403–412, 2012.

or they are still the pictures of the same object, but taken from different perspectives. By analyzing a group (called "window") of samples, the classification system may decide that the incoming 2-D images do not indicate that the classified 3-D object has been replaced by a new one. The fact that the changes are minor and follow an organized trend makes the classification task relatively easy by the means of detection and adaptation. When the class changes (i.e., the classified 3-D object is instantly replaced with a different one), then the stream of the 2-D images changes more drastically, therefore the gradual hidden changes in the feature vectors describing the consequent data samples are unlikely to be mistaken with the actual class change.

1.2 Sudden Concept Drift

Let us consider the same example, but with the sudden concept drift. As the concept shift usually does not follow an organized trend or sustain stable impetuosity, the changes in the rotation angle of the 3-D object are more abrupt and less organized, causing the following 2-D images to differ more from each other. Also, if the order of the 3-D objects presented to the classification system is chaotic (the classes change randomly), then the constantly changing 2-D images may indicate either a change of the class a hidden change in the data characteristics. The classification system then has to decide whether the incoming 2-D images are representing a new 3-D object or the same object, however seen from a different perspective.

Popular approaches to classify data with concept shift may be described in general by the following procedure [6]:

```
1. if a batch of samples indicates concept shift,
2. then, use an outdated classifier and train a new classifier
   when data with class labels are available,
3. else, use the trained classifier.
```

Fig. 1. Pseudo-code of the algorithm classifying data stream with concept shift

An efficient approach to deal with the concept shift is by using classifier ensembles, which have proven to achieve better overall classification accuracy than the single classifiers in the dynamically changing environments [18]. Classifier ensembles perform the classification on the basis of mutual decisions, achieved by the majority or weighted voting [3], with weights distributed in various manners, depending on a classification system and a scenario.

1.3 Recurring Concept Shift

Recurring concept shift is a special case of concept shift. In this model the previous concepts may reappear in the data stream. In the example with 2-D images of 3-D objects it would mean that the rotation angles of some 3-D objects could repeat. Recurring concept shift has been introduced by Widmer *et al.* [19]. Authors proposed to deal with this phenomenon on the basis of the repository of known concepts with the corresponding classifiers, which are kept in the memory to be used for the classification of the future data if the concept recurs [12][5][9].

The pseudo-code of algorithm which deals with the recurring concept shift is presented in Fig.2.

```
1. if a batch of samples indicates concept shift,
2. then if the concept is known (e.g. it has occurred in the
   past),
     a) then use a classifier trained on the historical concept,
     b) else use an outdated classifier and train a new
        classifier when the data with class labels are
        available.
2. else, use the trained classifier.
```

Fig. 2. Pseudo-code of classification algorithms for the recurring concept shift data streams

The only difference between the procedure presented above and the one described in Fig.1 is in point 2, in which the classification system reuses the already trained classifier to classify the new data after the shift.

An important point in both procedures is the first step, which requires a drift detection mechanism. Most algorithms described in literature detect the concept drift in data stream on the basis of classification error-rate fluctuations [15][10], which can only be acquired when the knowledge of the true class labels of data samples is available. Such assumption does not adhere to the definition of intelligent classification systems, which should be self-sufficient and as independent from the knowledge of the data labels as possible. Although the drift detection mechanism is important for the classification systems, it can be considered as a separate research area and it is not implemented in the presented solution.

2 Simulated Recurrence

Recurrence of concept allows the classification system to perform the classification with the use of the classifiers previously trained on the recurring concept, lowering the costs of wrong decisions and minimizing the need for access to true class labels, which is normally required to train a new classifier with the new concept data. We aim to implement this procedure in the non-recurring concept shift scenario by simulating the recurrence of the concepts.

Assuming that the degree of the concept shift is limited i.e., changes of expected values of conditional probability density functions are finite in size, the system trains the classifiers on the data artificially generated from each simulated concept within the allowed concepts area and keeps them stored in the classifier repository for later use. The classifiers are grouped into ensembles, from which the most appropriate one is chosen to perform the classification of the new data window drawn from the data stream.

If the base classifier, trained on the real data is denoted as Ψ, then the 'artificial classifiers', trained on the data samples generated from simulated concepts are described as $\Psi^{(1)}$, $\Psi^{(2)}$, ..., $\Psi^{(n)}$, where n is the number of simulated concepts.

The pseudo-code of the proposed solution is presented in Fig. 3.

```
Parameters: $C$: current data distribution (concept),
    $D_C$: set of data distributed according to $C$,
    $\Psi_C$: classifier trained on the dataset $D_C$,
    $SC$: simulated concept,
    $D_{SC}$: set of data generated according to $SC$,
    $\Psi_{SC}$: classifier trained on generated dataset $D_{SC}$,
    $E_\Psi$:ensemble of classifiers,
    $R_\Psi$: repository of classifiers,
    $B$: batch of samples drawn from the data stream.
  // Simulating concepts and training classifiers
for $i$ = 1:number of classifiers in repository do
  create new $SC_i$ on the basis of $C$,
  generate $D_{SCi}$ from $SC_i$,
  train $\Psi_{SCi}$ with $D_{SCi}$,
  add $\Psi_{SCi}$ to $R_\Psi$.
end for
Find_the_best_ensemble(B)
```

Fig. 3. Pseudo-code of the data stream classification with the simulated recurrence approach

The simulated recurrence method [17] can be explained more clearly on the basis of the previously used example with the 3-D objects represented by the 2-D images. As before, let's assume that the current concept of a 3-D object is represented by the 2-D images of the certain perspective of this object. Shift impetuosity is limited by the rotation of the 3-D object, e.g. at most 30 degrees in each direction, creating an area of possible concepts. Simulated recurrence is implemented by artificially generating the 2-D images of various perspectives of the 3-D objects within the allowed concepts area and with each perspective a separate classifier is trained and stored in the repository. Both the shift impetuosity limit and the number of 2-D images generated this way are defined as the preliminary parameters for the algorithm.

3 The Learning Algorithm Overview

The aim of the learning algorithm is to find the best ensemble for classification of the given window of data, what implies defining a quality measure for classifier ensemble. In our scenario, the most wanted ensemble is capable of classifying the most possible concepts accurately, what leads to an optimization task of finding a committee with the most diverse classifiers.

3.1 Measuring the Classifier Ensemble Diversity

Let us assume that we have n classifiers, $\Psi^{(1)}$, $\Psi^{(2)}$, ..., $\Psi^{(n)}$ at our disposal. Each classifier decides if an object belongs to a class $i \in \boldsymbol{M} = \{1, \dots, M\}$ on the basis of the observation$x = \left[x^{(1)}, x^{(2)}, \dots, x^{(d)}\right]^T \in X$.

We select classifiers from an available pool to the ensemble which is represented by the following binary word:

$$W_s = \left[w_s^{(1)}, w_s^{(2)}, \ldots, w_s^{(n)}\right], \tag{1}$$

where

$$w_s^{(k)} = \begin{cases} 1, if\ the\ k-th\ classifier\ belongs\ to\ ensemble \\ 0, otherwise \end{cases}$$

The combined classifier $\mathit{\Psi}_S$ makes decisions on the basis of the majority voting rule, denoted as:

$$\overline{\Psi_s}(x) = argmax_{j\in M} \sum_{l=1}^{n} \delta\left(j, \Psi^{(l)}(\mathrm{x})\right) w_s^{(l)} \tag{2}$$

where
$\delta(j,i) = \begin{cases} 0\ if\ i \neq j \\ 1\ if\ i = j \end{cases}.$

Let us focus on establishing an ensemble for the combined classifier. As we mentioned above, the aim is to find an ensemble model which assures the highest diversity of the chosen ensemble, denoted as $DIV(\overline{\Psi})$. The literature describes a number of methods designed to measure the classifier diversity [20][2][7][11].We propose to use pair-wise measure which is based on the soft and crisp label outputs of the classifiers presented in [4]

Let $\Psi_{soft}^{(l)}(x_k) = \left[\Psi_{x_k,1}^{(l)}, \Psi_{x_k,2}^{(l)}, \ldots, \Psi_{x_k,M}^{(l)}\right]$ be the soft label output vector of classifier $\Psi^{(l)}$ for observation x_k.

First, a so-called *certainty index, C* is calculated for each observation x_k belonging to a set of m observations $x_{1,\ldots,m} \in X$, describing the degree of the $\Psi^{(l)}$ classifier's confidence that observation x_k belongs to a class i:

$$C_{x_k}^{(l)} = \Psi_{x_k,i}^{(l)} - \frac{1}{M-1} \sum_{j\in M/\{i\}} \Psi_{x_k,j}^{(l)}\ , k = 1, \ldots, m\,, \tag{3}$$

where

$$\Psi_{x_k,i}^{(l)} = \max\left\{\Psi_{x_k,1}^{(l)} \Psi_{x_k,2}^{(l)}, , \ldots, \Psi_{x_k,M}^{(l)}\right\}. \tag{4}$$

Next, a set of m observations $x_{1,\ldots,m} \in X$ is divided into sets X_1 and X_0, representing observations for which the measured pair of classifiers $\Psi^{(l)}$ and $\Psi^{(h)}$ agrees or disagrees on the class affiliations, respectively. Two parameters are calculated on the basis of the observations belonging to sets X_1 and X_0:

$$A_{l,h} = \sum_{x_k \in X_1} \left| C_{x_k}^{(l)} - C_{x_k}^{(h)} \right| , \qquad D_{l,h} = \frac{1}{2} \sum_{x_k \in X_0} \left| C_{x_k}^{(l)} + C_{x_k}^{(h)} \right| . \tag{5}$$

The measure of diversity between a pair of classifiers $\Psi^{(l)}$ and $\Psi^{(h)}$ is then calculated as follows:

$$DIV\left(\Psi^{(l)}, \Psi^{(h)}\right) = \frac{A_{l,h} + D_{l,h}}{m} . \tag{6}$$

The cumulative diversity of classifier ensemble Ψ_s is a sum of diversities between all pairs of the classifiers in the ensemble:

$$DIV(\overline{\Psi_s}) = \sum_{l=1}^{n} \sum_{h=l}^{n} DIV\left(\Psi^{(l)}, \Psi^{(h)}\right) w_s^{(l)} w_s^{(h)} . \tag{7}$$

3.2 Genetic Approach for Creating the Classifier Ensemble

There are plethora of algorithms, which can be used to solve the mentioned optimization task, one example is a genetic algorithm [13]. With properly adjusted model of chromosomes, fitness function and the mutation and crossover operators it can serve as a base for our solution.

The genetic algorithms have already proven to be useful for solving similar problems, which can be found in a number of positions in the machine learning literature. An interesting approach can be found in [8], where the weights assigned to the elementary classifiers are the real numbers. This idea has inspired our research by representing the chromosome as the vector W_s, which is a model of the parameters of an ensemble.

Each chromosome corresponds to a given realization of the compound classifier, the quality of which can be obtained by computing $DIV(\overline{\Psi})$ according to (7). Referring to the naming convention used in the evolutionary computation literature, the quality of a chromosome is described by the fitness function, denoted as:

$$\Phi(W_s) = DIV(\overline{\Psi_s}) . \tag{8}$$

Initially, a set of members considering all the constraints of the model is generated. A value of the fitness function is calculated according to (8) for each member in the population, on the basis of the ensemble diversity measure (7). A certain number of members which are characterized by the highest fitness function value are drawn from the population as the elite. The elite is put directly into the descendant population, not being treated neither by the mutation or crossover operators nor by the selection procedure.

The mutation is applied with a certain probability to every other member in the population by inverting a randomly chosen bit value in the individual's W_s vector. As a result, the classifiers previously not taken into consideration have a chance of being selected to the ensemble.

The crossover operator generates two offsprings from two parents according to the traditional crossover rule, namely by partially swapping the W_s vectors between the parents. The crossover model assumes, that if the crossover takes place for a group of individuals, then a pair is chosen randomly as parents and 1/3 of the bits in the W_s vector of one parent are randomly selected and exchanged with the corresponding bits of the W_s vector of the second parent, creating two new W_s vectors, namely the children.

Selection of the individuals for the new population is performed from the merged descendant population with the set of the individuals created by the mutation and crossover operators. The probability of selecting a particular individual is proportional to the value of its fitness function value. The number of members in the new population stays the same as in the previous population, including the elite which was previously promoted.

The optimization process ends when the results obtained by the best member deteriorate in the course of a given number of the subsequent learning cycles.

4 Experiments

The aim of experiments is to evaluate the influence of the number of the classifiers in the ensemble and the number of classifiers in the repository on the performance of the classification system with simulated recurrence.

4.1 Setup

The evaluation takes place on 100 windows of 40 samples drawn randomly from each of 78 possible concepts and the accuracy of the system is measured on each window by dividing the number of correctly classified samples by the total number of samples in the window. Parameters of the learning algorithm are set as follows:

- upper limit for the number of learning algorithm cycles: 30,
- population quantity: 20,
- elite fraction size: 50%,
- probability of crossover: 50%,
- probability of mutation: 20%.

The analysis of results is performed by comparing the mean accuracies achieved by the proposed classification system for 10, 20 and 50 classifiers in the repository and 4, 6 and 8 classifiers in the ensemble. The efficiency obtained by the single classifier trained only with the original data is also presented for the reference.

A simple linear quadratic classifier is used as a base classifier in the experiments. The experimental scenario is based on the benchmark dataset from the UCI Machine Learning Repository [1], the Wine Dataset. This dataset consists of 178 samples distributed non-uniformly between 3 classes (59, 71 and 48 samples for each class respectively), which are described by the vectors of 13 real numeric features. During the experiments, all the original dataset is used as an available training and reference data for the system. The artificial datasets are generated on the basis of the original dataset by simulating the possible shifts in the original concept and the classifiers repository is created by training classifiers on each of the artificial datasets.

To perform the analysis of our classification system, the classification environment needs to be influenced by the shifts in concept. As the reference dataset does not include the concept shift, this feature needs to be simulated.

Shift in the original dataset is simulated by rotating two features in all samples, leaving the same class labels. Let us assume, that the samples in the reference dataset are described by the vectors of features which take a general form of $x = [f_1, f_2, f_3, f_4]$. To simulate a shift in concept, we exchange the values of two features for each sample in the dataset, so for each $x \in X$ we swap the feature values, e.g.

Before shift: $x = [f_1, f_2, f_3, f_4]$,

After shift: $x' = [f_2, f_1, f_3, f_4]$.

In the experimental scenario, the samples are represented by 13 features, what results in 78 possible concepts in total.

4.2 Results

To validate the statistical significance of the results, the vectors of mean accuracies obtained by each ensemble for all possible concepts have been tested with a paired t-test to reject the null hypothesis with 5% probability. The cumulative results over all possible concepts are presented in the table below:

Table 1. Mean accuracies achieved by various configurations of the classification system on 100 windows of 40 samples for each of 78 possible concepts

No. of classifiers in repository	No. of classifiers in ensemble	**Achieved mean accuracy [%]**	
		By each ensemble size	**By best ensemble**
10	4	62,23	**62,23**
	6	60,19	
	8	58,34	
20	4	66,45	**66,45**
	6	64,85	
	8	62,05	
50	4	80,90	**80,90**
	6	78,22	
	8	75,55	
Single classifier		**48,85**	

Each possible concept is evaluated separately and the numbers of concepts with corresponding mean accuracies are grouped into 10 equal bins as the accuracy histograms – below the worst and the best scenario, respectively:

Inspired by the recurring concept drift, the method of implementing a simulated recurrence into a static classification system has proven to extend its ability to classify data with shifting concept, without an additional concept drift detection mechanism. Even with only 13% of all possible concepts simulated as recurring, the classifier ensemble algorithm performs 25% better overall than the single classifier trained solely on the reference dataset and achieves an impressive mean accuracy of above 80% over all the possible concepts if more concepts are simulated.

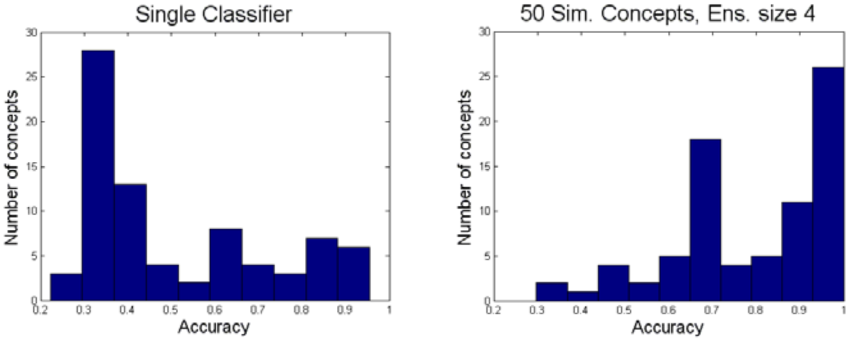


Fig. 4. Histograms representing the mean accuracies over all possible concepts of two system configurations (single classifier and an ensemble of 4 classifiers with the repository of 50 simulated concepts)

Surprising result is the lower accuracy for larger ensembles. The reason for this is when there are more than 4 classifiers in the ensemble it may lead to situations with more than one class label getting the most votes (e.g. 3-3-2 votes from 8 or 4-4-4 from 12 classifiers in the committee, in which case the label is chosen randomly).

5 Conclusions

The model of trained ensemble which is able to deal with concept shift problem was presented. The proposed method exploits strength of recurring concept drift model and adapts it to the problem of concept shift where the changes of the chosen probability characteristics are limited in each drift step. Our proposition was evaluated on the basis of computer experiments carried out on the basis of chosen dataset from UCI. The obtained results are promising but we realized that their scope was limited therefore our future research will be focused on the three areas, namely:

1. Extend the efficiency tests of the simulated recurrence ensemble classification algorithm by evaluating the method on other reference datasets and implementing different shift simulating rules,
2. Develop an efficient concept drift detection system based on the simulated recurrence, able to be implemented into an existing classification system as an autonomous detector.
3. Form a fair statistical measure to compare the algorithms used for classification of streaming data with concept drift.

Acknowledgment. This work is supported in part by The Ministry of Science and Higher Education under the grant which is realizing in the period 2010-2013.

References

1. Frank, A., Asuncion, A.: UCI Machine Learning Repository. University of California, School of Information and Computer Science, Irvine, CA (2010), `http://archive.ics.uci.edu/ml`

2. Demsar, J.: Statistical Comparisons of Classifiers over Multiple Data Sets. Journal of Machine Learning Research 7, 1–30 (2006)
3. Dietterich, T.G.: Ensemble Methods in Machine Learning. In: Kittler, J., Roli, F. (eds.) MCS 2000. LNCS, vol. 1857, pp. 1–15. Springer, Heidelberg (2000)
4. Fan, T.G., Zhu, Y., Chen, J.M.: A new measure of classifier diversity in multiple classifier system. ICMLC 1, 18–21 (2008)
5. Gama, J., Kosina, P.: Tracking Recurring Concepts with Meta-learners. In: Lopes, L.S., Lau, N., Mariano, P., Rocha, L.M. (eds.) EPIA 2009. LNCS, vol. 5816, pp. 423–434. Springer, Heidelberg (2009)
6. Gama, J., Medas, P., Castillo, G., Rodrigues, P.P.: Learning with Drift Detection. In: Bazzan, A.L.C., Labidi, S. (eds.) SBIA 2004. LNCS (LNAI), vol. 3171, pp. 286–295. Springer, Heidelberg (2004)
7. Ho, T.K., Hull, J.J., Srihari, S.N.: Decision Combination in Multiple Classifier Systems. IEEE Transactions on Pattern Analysis and Machine. Intelligence 16(1), 66–75 (1994)
8. Jackowski, K., Wozniak, M.: Algorithm of designing compound recognition system on the basis of combining classifiers with simultaneous splitting feature space into competence areas. Pattern Analysis and Applications 12, 415–425 (2009)
9. Katakis, I., Tsoumakas, G., Vlahavas, I.P.: An Ensemble of Classifiers for coping with Recurring Contexts in Data Streams. In: 18th European Conf. on Artificial Intelligence, Greece, pp. 763–764 (2008)
10. Klinkenberg, R., Joachims, T.: Detecting Concept Drift with Support Vector Machines. In: Proceedings of the Seventeenth International Conference on Machine Learning (ICML), pp. 487–494. Morgan Kaufmann, San Francisco (2000)
11. Kuncheva, L.I., Whitaker, C.J.: Measures of Diversity in Classifier Ensembles and Their Relationship with the Ensemble Accuracy. Machine Learning 51(2), 181–207 (2003)
12. Li, P., Wu, X., Hu, X.: Mining Recurring Concept Drifts with Limited Labeled Streaming Data. Journal of Machine Learning Research - Proceedings Track, 241–252 (2010)
13. Michalewicz, Z.: Genetics Algorithms + Data Structures = Evolutions Programs. Springer, Heidelberg (1996)
14. Narasimhamurthy, A.M., Kuncheva, L.I.: A framework for generating data to simulate changing environments. In: Proc. IASTED, Artificial Intelligence and Applications, Innsbruck, Austria, pp. 415–420 (2007)
15. Nishida, K., Yamauchi, K.: Learning, detecting, understanding, and predicting concept changes. In: Proc. of IJCNN, pp. 2280–2287 (2009)
16. Pechenizkiy, M., Bakker, J., Zliobaite, I., Ivannikov, A., Kärkkäinen, T.: Online mass flow prediction in CFB boilers with explicit detection of sudden concept drift. SIGKDD Explorations 11(2), 109–116 (2009)
17. Sobolewski, P., Woźniak, M.: Artificial Recurrence for Classification of Streaming Data with Concept Shift. In: Bouchachia, A. (ed.) ICAIS 2011. LNCS, vol. 6943, pp. 76–87. Springer, Heidelberg (2011)
18. Wang, H., Fan, W., Yu, P.S., Han, J.: Mining concept-drifting data streams using ensemble classifiers. In: Proc. of the 9th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 226–235. ACM Press, New York (2003)
19. Widmer, G., Kubat, M.: Learning in the Presence of Concept Drift and Hidden Contexts. Machine Learning 23(1), 69–101 (1996)
20. Windeatt, T.: Diversity measures for multiple classifier system analysis and design. Information Fusion 6(1), 21–36 (2005)

Neighborhood Selection and Eigenvalues for Embedding Data Complex in Low Dimension

Jiun-Wei Liou and Cheng-Yuan Liou

Department of Computer Science and Information Engineering
National Taiwan University
{r92001,cyliou}@csie.ntu.edu.tw

Abstract. LLE(Local linear embedding) and Isomap are widely used approaches for dimension reduction. For LLE, the neighborhood selection approach is an important research issue. For different types of datasets, we need different neighborhood selection approaches to have better chance for finding reasonable representation within the required number of dimensions. In this paper, the ϵ-distance approach and a modified version of k-nn method are introduced. For LLE and Isomap, the eigenvectors obtained from these methods are much more discussed, but there are more information hidden in the corresponding eigenvalues which can be used for finding embeddings contains more data information.

Keywords: Dimension reduction, Local linear embedding, Isomap, k-nearest neighbors, ϵ-distance.

1 Introduction

LLE[1] is a well known approach for representing the structure of high dimensional data within low dimensional embeddings. The LLE algorithm contains 3 steps as following:

1. Find neighborhood relation for every points in the dataset.
2. Solve for reconstruction weights.
3. Compute embedding coordinates using the solved weights.

The first step of LLE algorithm is to find out the neighborhoods of every points. Traditionally, the k-nearest neighbors approach is the most widely used one. This approach has many advantages such as easy to implement, suitable for most of cases, fast enough and can be parallelized and further accelerated[2]. But for some other type of dataset, the k-nn approach will face difficulties. For example, for non-uniform sampling datasets or datasets contains complex structure, the selection of k may be a serious problem. For these kind of problems, the ϵ-distance approach is introduced.

Although there are already attempts for using other neighborhood functions, such as weighted k-nn[3][4][5][6], clustering approaches[7], or including k-means[7][8]. But these alternatives are still based and analyzed on original

J.-S. Pan, S.-M. Chen, N.T. Nguyen (Eds.): ACIIDS 2012, Part I, LNAI 7196, pp. 413–422, 2012.

k-nn method. In this paper, the ϵ-distance will be taken into main consideration as a different conceptual method from k-nn for dealing with more complex datasets.

The last step of LLE requires computation for finding minimal eigenvalues and corresponding eigenvectors in a positive semi-definite matrix. Since the working precision for computer is always limited, directly finding eigenvalues nearest to zero will get eigenvalue near to machine epsilon which could be zero if we have infinite precision. The detail for this problem will be discussed later.

Isomap[9] is another popular approach for embedding data in low dimension. The Isomap algorithm also contains 3 steps as below:

1. Construct neighborhood graph.
2. Compute shortest paths.
3. Construct embedding coordinates using largest eigenvectors according to some computed matrix.

Since the k-nn and ϵ-distance approaches for neighborhood selection are already built in the Isomap approach, the focus for this paper is on comparison of eigenvalues obtained from LLE and Isomap. Although in [10], there are some brief analysis about the eigenvalues obtained from LLE, the analysis was focus on choosing the correct number of embedding dimensions. In this paper, we only map data to 2 embedding dimensions, so the choosing only occurs in finding appropriate k for k-nn and ϵ for ϵ-distance.

2 Method

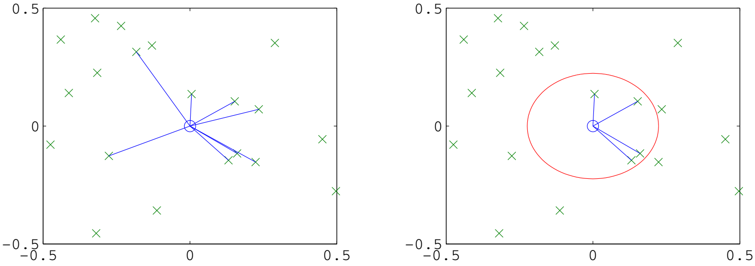
2.1 Neighborhood Selection

For neighborhood selection, we will use two approaches. The first method is k-nearest neighbors. The k-nearest neighbors method is finding k nearest data points for each data point in the dataset as its neighbors. For the ϵ-distance approach, a point is a neighbor of a certain point p if the Euclidean distance from the point to p is no more than a certain distance ϵ.

Neighborhood selection example for k-nn as 8-nn can be shown in Fig. 1(a), while example for ϵ-distance within radius $\epsilon = \sqrt{0.05}$ can be shown in Fig. 1(b).

2.2 K-nn Modification

In the k-nn method, the number of neighborhood for each point k should be a fixed integer, but this restriction is too strong so that the number of possible embeddings can be generated from k-nn is too small. For resolving this issue, a simple modification is proposed to perform the original k-nearest neighbors and then insert one more neighbor for some points with nearest $k + 1$-th neighbor. Since the Isomap method is only using for comparison of trends of eigenvalues, the k-nn modification is not used.



(a) 8-nn selection. (b) $\sqrt{0.05}$ selection.

Fig. 1. Different neighborhood selection approaches

2.3 Eigenvalues from LLE and Isomap

The last step of LLE requires finding d smallest eigenvalues which are not zero from a matrix computed from weights which is supposed to be positive semi-definte. The matrix to obtain d smallest eigenvalues created from ϵ-distance method, is expected to have more zero eigenvalues than the matrix created from k-nn because of imbalanced number of neighbors selected from ϵ-distance method. When performing eigendecomposition, the originally zero eigenvalues will be affected by machine error, so the original mechanism for finding d smallest eigenvalues will fail by choosing machine epsilons which originally should be considered as zero.

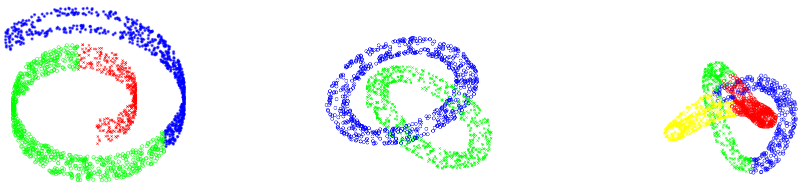
So the proposed modification is using the original eigenvalues searching program to find d eigenvalues nearest to some small value significantly larger than machine epsilon, and the small value can be automatically enlarged to escape from still finding eigenvalues within machine epsilon. Since the original Isomap algorithm needs d largest eigenvalues, the numerical problem occurred in LLE does not affect Isomap.

3 Experiment

3.1 Datasets

Some artificially generated datasets are used to test the ability of different approaches. The first dataset is from the swiss roll dataset with a hole to make sure all approaches are usable. The swiss roll dataset contains 2000 points as in Fig. 2(a). The second dataset is sampled from a dual circular tube dataset. The dataset samples one tube for 700 points, and samples the other tube for 800 points so that the sample size is summed up to 1500 points. The dataset can be shown in Fig. 2(b).

The third experimental data is sampled non-uniformly from a knot tube dataset. The sample size is 2000. The sampled data can be shown in Fig. 2(c). The fourth dataset is a database of 8-bit grey-level face images from the same person with different angles and moods which can be considered as true data. The number of images is 1965 and the resolution of images is 28 × 20.



(a) The swiss roll dataset. (b) The dual tube dataset. (c) The knot dataset.

Fig. 2. Simulation datasets for experiment

3.2 Parameters

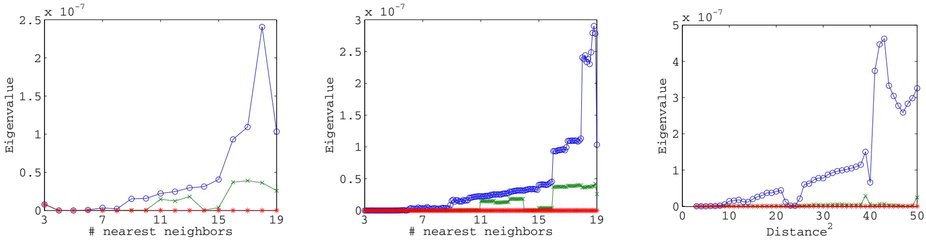
The LLE parameters can be separated as regularization parameters, eigenvalue solving parameters, and k-nn or ϵ-distance parameters. The regularization for all method introduced in previous section is the default parameter as 10^{-3} for simulation dataset. For the true dataset, the data dimension is 560, which should be larger than effective parameter range of k-nn approach, so the regularization for the true dataset is set to zero. For ϵ-distance on true dataset, the regularization is set to 10^{-3} for searching in a big enough range to ensure that no data is isolated. The eigenvalue solving parameter is only effective for ϵ-distance which indicates minimum non-epsilon eigenvalue is 10^{-14} for simulation dataset, and 10^{-12} for the true data. For Isomap, only the k in k-nn and the ϵ in ϵ-distance are considered as effective parameters. The detail for solving Isomap matrix will not be discussed in this paper.

3.3 Result

For the swiss roll with hole dataset, the corresponding eigenvalues solved from LLE for effective range of k or ϵ can be shown in Fig. 3(a) to Fig. 3(c). For LLE, the third eigenvalue in the figures, shown as red line for representative, should be always near zero and the corresponding eigenvector will never be used for the final embedding result. The reasonable embedding results for k-nn can be obtained when k is from 6 to 16. For k value between 11 and 13, since the second eigenvalue goes up, the embedding results enclose the hole in the swiss roll and are considered as descent embedding results. The selected embedding results for k-nn can be seen in Fig. 4. For the fractional nearest neighbors, we can obtain enhanced resolution for eigenvalues trends while the corresponding eigenvalues changes are similar to the original k-nn. The corresponding eigenvalue trends can be seen in Fig. 3(b).

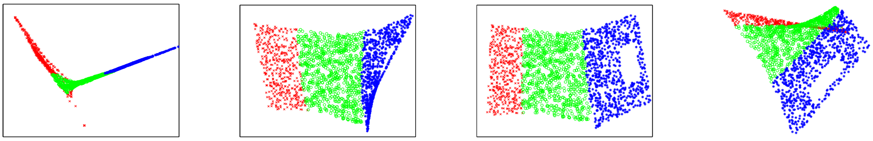
For the ϵ-distance approach, the eigenvalues for different parameters can be shown in Fig. 3(c). The effective ϵ range for the swiss roll with hole dataset is from $\sqrt{11}$ to $\sqrt{38}$, exclude $\sqrt{22}$ to $\sqrt{25}$ according to the figure. But the embedding result using ϵ within $\sqrt{22}$ to $\sqrt{25}$ get twisted but still reasonable embedding result, which is still worse than embbedding results from other reasonable ϵs. Some selected embedding results for ϵ-distance can be shown in Fig. 5.

Since LLE selected the d eigenvectors with least positive eigenvalues except zeros, the corresponding eigenvalues will generally increasing as the increasing parameters. The sharper change points indicate the major change of embedding results, such as from too less information to reasonable embeddings or from reasonable embeddings to complex structure which is no more understandable. The range between two change points will obtain similar embedding results.



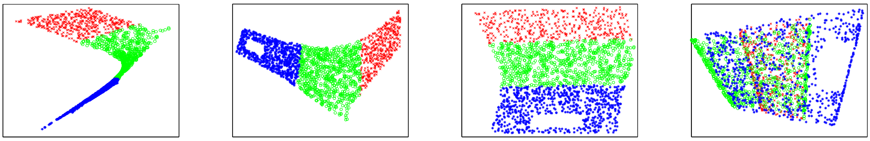
(a) k-nn (b) Fractional k-nn (c) ϵ-distance

Fig. 3. Eigenvalues from LLE on swissroll with hole dataset



(a) 6-nn (b) 12-nn (c) 14-nn (d) 16-nn

Fig. 4. Embedding results from LLE using k-nn on swissroll with hole dataset

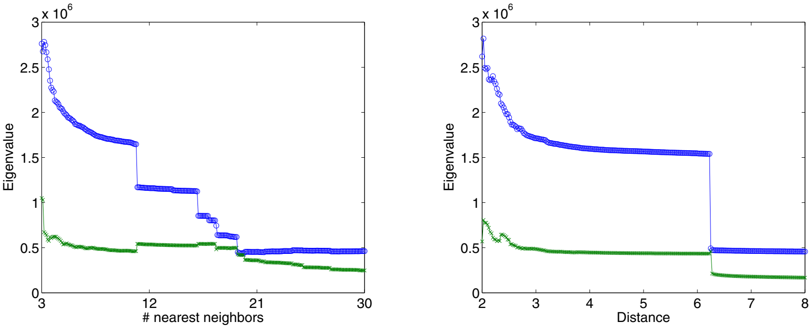


(a) $\epsilon = \sqrt{10.5}$ (b) $\epsilon = \sqrt{24}$ (c) $\epsilon = \sqrt{32}$ (d) $\epsilon = \sqrt{40}$

Fig. 5. Embedding results from LLE using ϵ-distance on swissroll with hole dataset

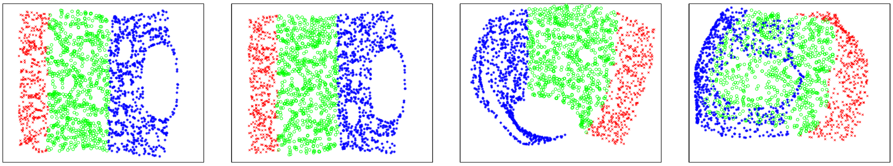
For Isomap applied on the swiss roll with hole dataset, the corresponding eigenvalue changes for k-nn and ϵ-distance methods can be shown in Fig. 6(a) and Fig. 6(b). For k-nn approach, the dramatic eigenvalue change from 8-nn to 10-nn only has implicit meaning for the embedding process which becomes significantly shown after 12-nn, and the embedding result starts to be folded after

16-nn. For ϵ-distance approach, the usable embedding range is approximately $2.5 < \epsilon < 6.25$. When $\epsilon \leq 6.25$, the embedding result is also folded and the hole can never be shown. Some selected results from Isomap using k-nn and ϵ-distance can be shown in Fig. 7 and Fig. 8. Since the page is limited, the corresponding embedding results will be omitted for other datasets.



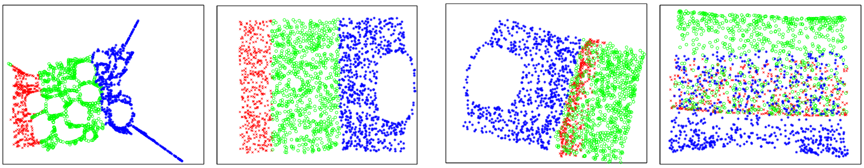
(a) k-nn (b) ϵ-distance

Fig. 6. Eigenvalues from Isomap on swissroll with hole dataset



(a) 8-nn (b) 10-nn (c) 12-nn (d) 18-nn

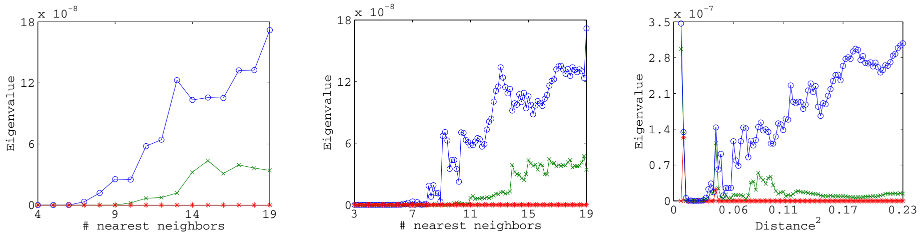
Fig. 7. Embedding results from Isomap using k-nn on swissroll with hole dataset



(a) $\epsilon = 2.4$ (b) $\epsilon = 5$ (c) $\epsilon = 6.25$ (d) $\epsilon = 6.3$

Fig. 8. Embedding results from Isomap using ϵ-distance on swissroll with hole dataset

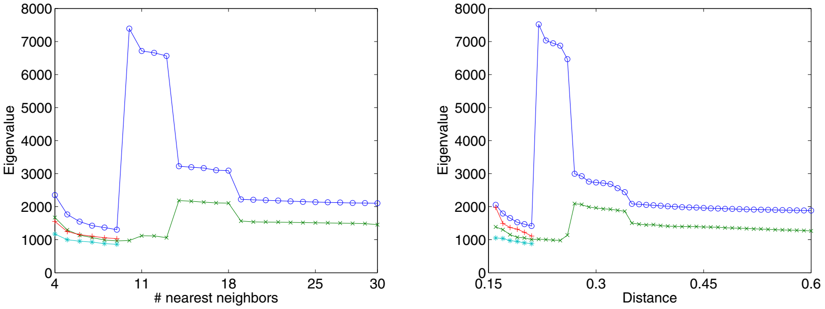
For the dual tubes dataset, the eigenvalues solved from LLE by different parameters using k-nn, fractional k-nn and ϵ-distance can be shown in Fig. 9. The effective range for k-nn is $10 \leq k \leq 14$. If $k > 14$, the two tubes will be intersected. The effective range for ϵ-distance is $\sqrt{0.05} \leq \epsilon \leq \sqrt{0.07}$. The significant change of the second eigenvalue determines the effective range in this dataset.



(a) k-nn (b) Fractional k-nn (c) ϵ-distance

Fig. 9. Eigenvalues from LLE on dual tubes dataset

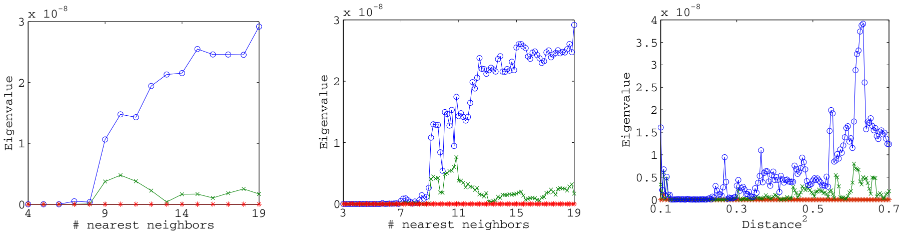
The eigenvalues solved by Isomap using k-nn and ϵ-distance can be shown in Fig. 10. The red and blue line shown the eigenvalues of the second component extracted from Isomap. Since the dual tube is not intersected within 3-D space, the separated dual tube result is more desired sometimes. The effective range for k-nn of Isomap is $k < 19$ while the effective range for ϵ-distance is $\epsilon < 0.35$. The different level of eigenvalues mapped to different embedding results from two tubes to dual tubes in the embedding space to twisted dual tubes.



(a) k-nn (b) ϵ-distance

Fig. 10. Eigenvalues from Isomap on dual tubes dataset

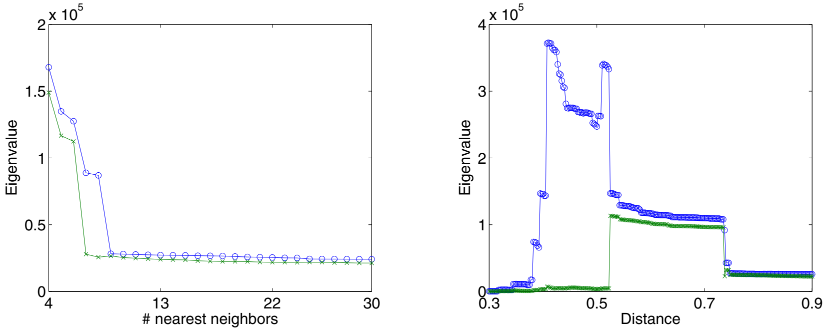
For the knot dataset, the eigenvalues solved from LLE by different parameters using k-nn, fractional k-nn and ϵ-distance can be shown in Fig. 11. Only 7-nn can successfully unfold the knot, while effective ϵ range is from $\sqrt{0.36}$ to $\sqrt{0.54}$ except $\sqrt{0.46}$ to $\sqrt{0.48}$. The first eigenvalue change is more important in this dataset using LLE.



(a) k-nn (b) Fractional k-nn (c) ϵ-distance

Fig. 11. Eigenvalues from LLE on knot dataset

The eigenvalues solved by Isomap using k-nn and ϵ-distance can be shown in Fig. 12. The effective k is $4 \leq k \leq 8$, while $6 \leq k \leq 8$ can only obtain descent twisted tube which is not intersected. The effective ϵ is $0.525 \leq \epsilon \leq 0.735$.

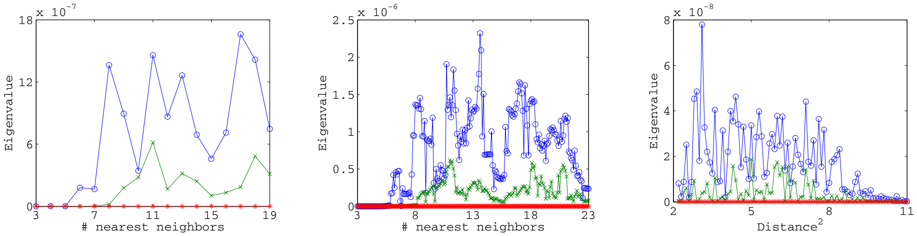


(a) k-nn (b) ϵ-distance

Fig. 12. Eigenvalues from Isomap on knot dataset

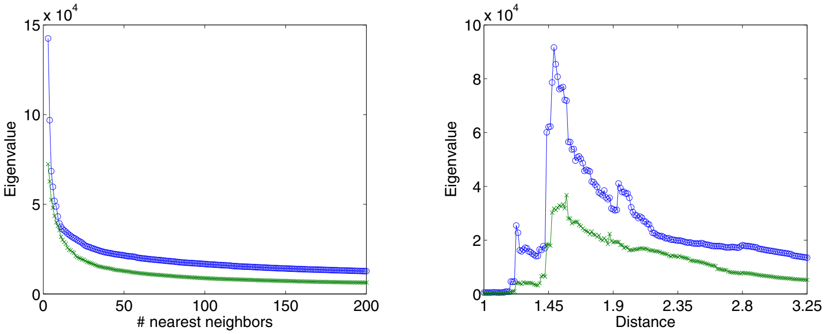
For the face dataset, the eigenvalues solved by LLE using the three approaches can be shown in Fig. 13. The true data is much more complex, so the strategy for finding suitable embbeding is no more applicable for the true data. For ϵ-distance approach, the trade-off between enough number of points included and not too many connections included becomes serious decision because some data are far away from other data and will almost always be excluded from the final embedding result.

The eigenvalues solved by Isomap using k-nn and ϵ-distance can be shown in Fig. 14. For Isomap, similar embeddings can be obtained from k-nn and ϵ-distance from the corrsponding value regions. The change point for Isomap using ϵ-distance show more changes from different distance range while using k-nn, we cannot find any significant change point. The first jump for distance close to 1.45 indicates the information from insufficient to barely enough to build up an embedding. The second jump means including another isolated large component so that the first eigenvalue jumps up again just as what happened in the dual tube dataset. The distance between flat area will just give similar embedding results.



(a) k-nn (b) Fractional k-nn (c) ϵ-distance

Fig. 13. Eigenvalues from LLE on face dataset



(a) k-nn (b) ϵ-distance

Fig. 14. Eigenvalues from Isomap on face dataset

4 Conclusion

From the eigenvalues obtained from LLE and Isomap, we can observe the evolution of unfolding internal data structure across different parameters. Using too small distance or number of neighbors will result in insufficient connections so that the global view of data is considered as incomplete. While using too large distance or number of neighbors will force LLE and Isomap to reserve too much information to represent within the corresponding number of embedding dimensions. The eigenvalues obtained from different parameters are proven to be a possible criteria to choose relatively good distance or number of neighbors for good enough embeddings within the required number of embedding dimensions. But for real world dataset, the searching range for parameters is data dependent and should be large enough to ensure containing better embedding results, although too large search range will need too much time to finish. The resonable search range for the k-nn or ϵ-distance approach for embedding dataset in fixed number of embedding dimensions is still remained as a future research issue.

Acknowledgement. This work was supported by National Science Council under project number NSC 100-2221-E-002-234-MY3.

References

1. Sam, R., Lawrence, S.: Nonlinear Dimensionality Reduction by Locally Linear Embedding. Science 290(5500), 2323–2326 (2000)
2. Yeh, T.T., Chen, T.-Y., Chen, Y.-C., Shih, W.-K.: Efficient Parallel Algorithm for Nonlinear Dimensionality Reduction on GPU. In: IEEE International Conference on Granular Computing, pp. 592–597. IEEE Computer Society (2010)
3. Chang, H., Yeung, D.-Y.: Robust Locally Linear Embedding. Pattern Recognition 39, 1053–1065 (2006)
4. Pan, Y., Ge, S.S., Mamun, A.A.: Weighted Locally Linear Embedding for Dimension Reduction. Pattern Recognition 42, 798–811 (2009)
5. Wen, G., Jiang, L., Wen, J.: Local Relative Transformation with Application to Isometric Embedding. Pattern Recognition Letters 30, 203–211 (2009)
6. Zuo, W., Zhang, D., Wang, K.: On Kernel Difference-weighted K-nearest Neighbor Classification. Pattern Analysis and Applications 11, 247–257 (2008)
7. Wen, G., Jiang, L.-J., Wen, J., Shadbolt, N.R.: Clustering-Based Nonlinear Dimensionality Reduction on Manifold. In: Yang, Q., Webb, G. (eds.) PRICAI 2006. LNCS (LNAI), vol. 4099, pp. 444–453. Springer, Heidelberg (2006)
8. Wei, L., Zeng, W., Wang, H.: K-means Clustering with Manifold. In: Seventh International Conference on Fuzzy Systems and Knowledge Discovery, pp. 2095–2099. IEEE Xplore Digital Library and EI Compendex (2010)
9. Joshua, T., de Vin, S., John, C.L.: A Global Geometric Framework for Nonlinear Dimensionality Reduction. Science 290(5500), 2319–2323 (2000)
10. Lawrence, S., Sam, R.: Think Globally, Fit Locally: Unsupervised Learning of Low Dimensional Manifolds. Journal of Machine Learning Research 4, 119–155 (2003)

A Bit-Chain Based Algorithm for Problem of Attribute Reduction

Thanh-Trung Nguyen, Viet-Long Huu Nguyen, and Phi-Khu Nguyen

Department of Computer Science, University of Information Technology,
Vietnam National University HCM City, Vietnam
nguyen_thanh_trung_key@yahoo.com.vn,
nguyenhuuvietlong@gmail.com,
khunp@uit.edu.vn

Abstract. Rough set is a widespread concept in computer science and is applicable in many fields such as artificial intelligence, expert systems, data mining, pattern recognition and decision support systems. One of key problems of knowledge acquisition in theoretical study of rough sets is attribute reduction. Attribute reduction also called feature selection eliminates superfluous attributes in the information system and improves efficiency of data analysis process. But reducing attributes is a NP-hard problem. Recently, to overcome the technical difficulty, there are a lot of research on new approaches such as maximal tolerance classification (Fang Yang et al. 2010), genetic algorithm (N. Ravi Shankar et al. 2010), topology and measure of significance of attributes (P.G. JansiRani and R. Bhaskaran 2010), soft set (Tutut Herawan et al. 2010), positive approximation (Yuhua Qian et al. 2010), dynamic programming (Walid Moudani et al. 2010). However, there are still some challenging research issues that time consumption is still hard problem in attribute reduction. This paper introduces a new approach with a model presented with definitions, theorems, operations. Set of maximal random prior forms is put forward as an effective way for attribute reduction. The algorithm for seeking maximal random prior set are proposed with linear complexity, contributes to solve absolutely problems in attribute reduction and significantly improve the speed of calculation and data analysis.

Keywords: attribute reduction, maximal random prior set, rough set.

1 Introduction

Most of recent research on attribute reduction fall into four categories: the reduction based on discernibility matrix, the reduction based on positive region, the reduction based on random strategy and the reduction on special information systems.

The discernibility matrix is used to compute reductions and the core of dataset [20]. Simplifying the discernibility function retrieved from discernibility matrix is the most basic method to reduce superfluous attributes. There are much research on discernibility function from simplicity such as Johnson's algorithm and Genetic Reducer [3] to complexity such as the Universal reduction [19]. Besides, some scholars focus on changing discernibility matrix [9][12][17].

J.-S. Pan, S.-M. Chen, N.T. Nguyen (Eds.): ACIIDS 2012, Part I, LNAI 7196, pp. 423–431, 2012.

The positive region measures the positive of attributes based on coefficients and reducing bad attributes. There are a lot of research based on this technique. A Greedy Reduction Algorithm based on Consistency [18] and Distance Measure Assisted Rough Set Attribute Reduction [11] are samples. According to another study, researchers divided the entire dataset into subsets called topology and computes measure of significance of attributes to reduce redundant attributes in an information system where a large amount of data have to be processed [4]. A theoretic framework based on rough set theory, called positive approximation, was also introduced [6]. In addition, there are many papers proposed coefficients reflecting the important degree of each attribute [10][15][16], especially entropy and fuzzy entropy.

The random strategy improves time efficiency and space efficiency, hard problems of attribute reduction algorithm. Genetic Algorithm is the typical algorithm of random strategy. Many reduction techniques are based on this algorithm [2][3][13]. Moreover, dynamic programming is a useful technique to improve attribute reduction problem. It permits to explore the optimal sets of significant attributes and reduces the process complexity [7].

In reality, most of information systems are incomplete and incorrect. As an indispensable result, maximal tolerance classification was invented to classify incomplete information systems and reduce superfluous attributes [1]. One of other special information systems used to solve the incomplete and incorrect status is object oriented information system. Besides, researchers used an alternative approach based on AND and OR operations in multi-soft sets [5]. To contribute to treatments for the problem, fuzzy-rough reductions or fuzzy-rough feature selection is created and also a noticeably concept. Fuzzy entropy is applied to reduce attributes in this fuzzy-rough set [15].

In addition, there are still some special methods to reduce number of attributes such as Binary Conversion [8] or concept of lattice, an efficient tool for knowledge representation and knowledge discovery [14].

All contributions will be welcome, however small. According to that point of view, this paper introduces a new approach to simplify the discernibility function. Mathematical model on binary strings is presented and the maximal random prior forms set of binary strings is defined as a basis for finding a reduction form of the discernibility function.

2 Formulation Model

Definition 1 (bit-chain): $< a_1a_2 \ldots a_m >$ (for $a_i \in \{0,1\}$) is a m-bit-chain. Zero chain is a bit-chain with each bit equals 0.

Definition 2 (intersection operation $\overline{\cap}$***):*** The intersection operation $\overline{\cap}$ is a dyadic operation in bit-chains space.

$< a_1a_2 \ldots a_m > \overline{\cap} < b_1b_2 \ldots b_m > = < c_1c_2 \ldots c_m >$, $a_i, b_i \in \{0,1\}$, $c_i = \min(a_i,b_i)$

Definition 3 (cover operation $\hookrightarrow$***):*** A bit-chain A is said to cover a bit-chain B if and only if with every position having bit-1 turned on in B, A has a corresponding bit-1 turned on.

Let $A = < a_1a_2 \ldots a_m >$, $B = < b_1b_2 \ldots b_m >$, $(\forall b_{i=1..m} \mid (b_i = 1) \rightarrow (a_i = 1)) \Rightarrow A \hookrightarrow B$

Consequence 1: A bit-chain being the result of an intersection operation and differing from zero chain is always covered by two bit-chains generating it.

$(A \overline{\cap} B = C) \wedge (C \neq 0) \Rightarrow (A \hookrightarrow C) \wedge (B \hookrightarrow C)$

Definition 4 (maximal random prior form $\delta - S$): The maximal random prior form of a set S of bit-chains, denoted by $\delta - S$, is a bit-chain satisfing four criteria:

- Being covered most by elements in S.
- Being covered by the first element in S.
- Having number of bit 1 turned on as much as possible.
- If there are more than one bit-chain meeting three criteria above, the bit-chain chose to be the maximal random prior form of S is one covered by first elements in S.

For example, consider a set of 4-bit-chain <abcd>:

$$S = \{ \quad \begin{array}{cccccc} & a & b & c & d & \\ (& \boxed{1} & 0 & \boxed{1} & 1 &); \\ (& 0 & 0 & \boxed{1} & 1 &); \\ (& \boxed{1} & 1 & 0 & 0 &); \\ (& \boxed{1} & 0 & \boxed{1} & 0 &) \end{array} \quad \}$$

Review three bit-chains:

<0011>: has two bit-1 turned on but is only covered by two first bit-chains of S.

<1000>: has one bit-1 turned on and is covered by three bit-chains of S.

<0010>: has one bit-1 turned on and is covered by three bit-chains of S.

Between <1000> and <0010>, <0010> is covered by two first elements in S, so $\delta -$ S has to be <0010>.

Definition 5 (maximal random prior elements): Maximal random prior elements of set S of bit-chains have the following characteristics:

The first element (p_1) is form $\delta - S$

The second element (p_2) is form $\delta - S \backslash \{x \in S \mid x \hookrightarrow p_1\}$

The third element (p_3) is form $\delta - S \backslash (\{x \in S \mid x \hookrightarrow p_1\} \cup \{x \in S \mid x \hookrightarrow p_2\})$

…

The k^{th} element (p_k) is form $\delta - S \backslash (\{x \in S \mid x \hookrightarrow p_1\} \cup \{x \in \mathrm{S} \mid x \hookrightarrow p_2\} \cup ... \cup \{x \in S \mid x \hookrightarrow p_{k-1}\})$

and $S = \{x \in S \mid x \hookrightarrow p_1\} \cup \{x \in S \mid x \hookrightarrow p_2\} \cup ... \cup \{x \in S \mid x \hookrightarrow p_k\}$

Definition 6 (maximal random prior set): A set P containing all maximal random prior elements of a set S of bit-chains is called maximal random prior set of S.

Consequence 2: All elements in maximal random prior set P do not have any the same position where bit-1 turned on.

Consequence 3: When the bit-chains set is arranged in different orders, it will produce different maximal random prior sets.

Theorem 1: When the intersection operations are made between an element in S and elements in P, the results differing from zero chain will not cover each other.

Proof: According to *Consequence 2*, the results made from intersection operations of an element in S and elements in P will not have bit-1 turned on at the same position. So that, the bit-chains being results will not cover each other.

3 Algorithm for Finding Maximal Random Prior Set

3.1 Idea

Consider a Boolean function f is the intersection ($\wedge$) of n propositions. Each proposition in f is a union ($\vee$) of m variables a_1, a_2, ..., a_m. According to commutative law of Boolean algebra, n propositions of f can be changed into the form: $f = A_1 \wedge A_2 \wedge \ldots \wedge A_m$, with:

$A_1 = \wedge_{k_1} (a_1 \vee \ldots)$

$A_2 = \wedge_{k_2} (a_2 \vee \ldots)$ A_2 does not contain a_1

$A_3 = \wedge_{k_3} (a_3 \vee \ldots)$ A_3 does not contain a_1, a_2

…

$A_m = \wedge_{k_m} (a_m \vee \ldots)$ A_m does not contain $a_1, a_2, \ldots, a_{m-1}$

$\forall i = 1..m;\ 0 \le k_i \le n \mid k_1 + k_2 + \ldots + k_m = n;$

$\forall k_p \neq 0;\ 1 \le p \le m \mid A_p = a_p \vee X_p$; X_p is a certain proposition.

So, $f = \wedge A_p = \wedge (a_p \vee X_p) = (\wedge a_p) \vee (\wedge X_p)$

Clearly, $(\wedge a_p)$ is a reduction of f.

If n propositions in f are transformed into a set S of m-bit-chains, the maximal random prior set P will be a reduction of f.

According to the above analysis, an algorithm is took shape to construct maximal random prior set P of the bit-chains set S with the following main ideas:

Each element in set S will be inspected with the existing order in S. At the same time, the set P will be also created or modified correspondingly with the number of elements inspected in S.

The initial set P is empty. Obviously, the set S with one first element has the corresponding set P also containing only this first element.

Scanning the next element of S, the intersection operations made between this element and the existing elements of P to find out the new maximal random prior forms. If the new form is generated, it will replace the old form in P because this new form is covered by elements of S more than the old form, evidently. If the new form is not generated, obviously, the next element of S is one new maximal random prior form.

However, a question maybe be brought out. Whenever the next element in S is inspected, the element have to carry out intersection operations with the existing elements in P; at that time, we have two element groups listed such as: (1) the old elements of P, (2) the new elements created by the intersection operations. Maybe, the new elements will cover together or cover the old elements or be covered by the old elements. Therefore, whether the set P is ensured the consistency as *Consequence 2* stated? The answer is “Yes” since *Consequence 1* and *Theorem 1* are generated to ensure this.

3.2 Proposed Algorithm

```
FIND_MaximalRandomPriorSet
Input: m-bit-chains set S
Output: maximal random prior set P
1.  P = ∅;
2.  for each s in S do
3.    flag = False;
4.    for each p in P do
5.      temp = s $\overline{\cap}$ p;
6.      if temp <> 0 then//temp differs from zero chain
7.        replace p in P by temp;
8.        flag = True;
9.        break;
10.     end if;
11.   end for;
12.   if flag = False then
13.     P = P ∪ {s};//s becomes ending element of P
14.   end if;
15. end for;
16. return P;
```

3.3 Accuracy of the Algorithm

Theorem 2: *FIND_MaximalRandomPriorSet* algorithm can find out the maximal random prior set *P* of a bit-chains set *S* with a given order.

Proof by Induction:

With number of elements in *S* is 1, the only element in *S* is also form $\delta - S$. According to the algorithm, the only element in *S* is inserted into *P*. Then, the only element in *P* satisfies the definition of maximal random prior set. Since, *Theorem 2* is correct when *S* has 1 element.

Assume that *Theorem 2* is correct when *S* has *k* elements. We need to prove *Theorem 2* is correct when *S* has *k* + 1 elements, too.

Because *Theorem 2* is correct when *S* has *k* elements, we have the set *P* contains all maximal random prior elements of this set *S*.

When *S* has *k* + 1 elements, it means the original set *S* having *k* elements are added a new element.

According to *FIND_MaximalRandomPriorSet* algorithm, we make intersection operations between elements in current *P* and the new $(k + 1)^{\text{th}}$ element denoted s_{k+1} in *S* (line 4 and line 5):

- If the result of the intersection operation between s_{k+1} and an element p_i in *P* differs from zero chain (line 6), this result is form $\delta - S \backslash (\{x \in S \mid x$ ⤷ $p_1\} \cup \{x \in \mathrm{S} \mid x$ ⤷ $p_2\} \cup \ldots \cup \{x \in S \mid x$ ⤷ $p_{i-1}\})$, with *S* has *k* + 1 elements. Replace p_i in *P* by this new result element (line 7). When s_{k+1}, together with p_i, create a new maximal random prior form, we terminate intersection operations between s_{k+1} and remaining elements in P (line 9).

- If all intersection operations between s_{k+1} and each element in P return zero chain, it means s_{k+1} does not cover any element in P. Thus, the element s_{k+1} is form $\delta - \{s_{k+1}\}$, then s_{k+1} is inserted into P (line 13).

In both cases, we receive the set P satisfing the properties of the maximal random prior set of S. So, *Theorem 2* is correct when S has $k + 1$ elements.

In conclusion: *FIND_MaximalRandomPriorSet* algorithm can find out the maximal random prior set P of a bit-chains set S with a given order.

4 Attribute Reduction in Rough Set

The maximal random prior set P is useful in solving and reducing Boolean algebra functions. One of the most important applications of the set P is finding out a solution of attribute reduction problem in rough set.

4.1 Rough Set

In rough set theory, information system is a pair (U, A), where U is a non-empty finite set of objects and A is a non-empty finite set of attributes. A decision system is any information system of the form $(U; A \cup \{d\})$, where $d \not\subset A$ is decision attribute.

Table 1. A decision system “Play Sport”

	Wind	Temperature	Humidity	Outlook	Play Sport
x_1	Strong	Hot	Normal	Sunny	Yes
x_2	Strong	Mild	Normal	Rain	No
x_3	Weak	Hot	Normal	Rain	No
x_4	Weak	Cool	High	Rain	Yes

With $|U|$ denotes cardinal of U, discernibility matrix of a decision system is a symmetric $|U|\text{x}|U|$ matrix with each entry $c_{ij} = \{a \in A \mid a(x_i) \neq a(x_j)\}$ if $d(x_i) \neq d(x_j)$, otherwise $c_{ij} = \varnothing$.

Table 2. Discernibility matrix of decision system “Play Sport”

	x_1	x_2	x_3	x_4
x_1	∅	∅	∅	∅
x_2	b,d	∅	∅	∅
x_3	a,d	∅	∅	∅
x_4	∅	a,b,c	b,c	∅

Table 2 presents a discernibility matrix of decision system “Play Sport” where a, b, c, d denote Wind, Temperature, Humidity and Outlook, respectively.

Discernibility function is a Boolean function retrieved from discernibility matrix and can be defined by the formula $f = \wedge \{ \vee c_{ij} \mid c_{ij} \neq \varnothing \}$. According to *Table 2*, we have discernibility function $f = (b \vee d) \wedge (a \vee d) \wedge (a \vee b \vee c) \wedge (b \vee c)$.

Discernibility function can be simplified by using laws of Boolean algebra. All constituents in the minimal disjunctive normal form of this function are all reductions of decision system [20]. However, simplifying discernibility function is a NP-hard problem and attribute reduction is always the key problem in rough set theory.

4.2 The Maximal Random Prior Set and Attribute Reduction Problem

Consider a discernibility function *f* retrieved from discernibility matrix of a decision system with *m* attributes has *n* constituents. Each constituent in this function will be transformed into an *m*-bit-chain, with each bit denotes an attribute. The function will be converted into a set *S* has *n* bit-chains. The maximal random prior set *P* of the set *S* is the simplification of discernibility function *f*.

Set *P* shows (some) reduction(s) of function *f*. With each bit-chain in *P*, the position where bit-1 is turned on need to be noticed. Value 1 of a bit means that the corresponding attribute will appear in reduction of *f*. The collection of all attributes retrieved from set *P* is a simplification of discernibility function *f*.

Example: According to discernibility function *f* of decision system in *Table 1*, the set *S* includes:

$$S = \{ \quad \begin{array}{l} (\,0\ 1\ 0\,\boxed{1}\,); \\ (\,1\ 0\ 0\,\boxed{1}\,); \\ (\,1\,\boxed{1\ 1}\,0\,); \\ (\,0\,\boxed{1\ 1}\,0\,) \end{array} \qquad \}$$

Initialize $P = \varnothing$. Scan all elements in *S*

S[1] = (0 1 0 1) → insert (0 1 0 1) into *P* → *P* = { (0 1 0 1) }

S[2] = (1 0 0 1) → (1 0 0 1) $\overline{\cap}$ (0 1 0 1) = (0 0 0 1) → replace (0 1 0 1) in *P* by (0 0 0 1) → P = { (0 0 0 1) }

S[3] = (1 1 1 0) → (1 1 1 0) $\overline{\cap}$ (0 0 0 1) = (0 0 0 0) → insert (1 1 1 0) into *P* → *P* = { (0 0 0 1); (1 1 1 0) }

S[4] = (0 1 1 0) → (0 1 1 0) $\overline{\cap}$ (0 0 0 1) = (0 0 0 0); (0 1 1 0) $\overline{\cap}$ (1 1 1 0) = (0 1 1 0) → replace (1 1 1 0) in *P* by (0 1 1 0) → P = { (0 0 0 1); (0 1 1 0) }

(0 0 0 1) → d and (0 1 1 0) → $b \vee c$

So, minimal function $f = d \wedge (b \vee c)$.

In conclusion, $(d \wedge b)$ and $(d \wedge c)$ are two reductions of discernibility function *f*.

5 Trial Installation

FIND_MaximalRandomPriorSet algorithm is developed and tested on a personal computer with specification: Windows 7 Ultimate 32-bit, Service Pack 1 Operating System; 4096MB RAM; Intel(R) Core(TM)2 Duo, E7400, 2.80GHz; 300GB HDD. Programming language is C#.NET on Visual Studio 2008. The results of some testing patterns:

Table 3. Some testing patterns of *FIND_MaximalRandomPriorSet* algorithm

Length of bit-chain	Number of bit-chains	Time (unit: second)
10	1,000,000	0.2184004
10	2,000,000	0.3900007
10	5,000,000	1.0140017
10	10,000,000	2.0436036
25	1,000,000	0.2808005
25	2,000,000	0.546001
25	5,000,000	1.123202
25	10,000,000	2.8236049
50	1,000,000	0.2496004
50	2,000,000	0.7176013
50	5,000,000	1.9032033
50	10,000,000	3.978007
60	1,000,000	0.3744007
60	2,000,000	0.7644014
60	5,000,000	1.9344034
60	10,000,000	4.1964073

6 Conclusion and Future Work

The experimental result reflects the efficiency and accuracy of *FIND_MaximalRandomPriorSet* algorithm. The complexity of this algorithm is $n.2^m$ where n is the number of bit-chains in the set S and m is the length of a bit-chain. In fact, m is often unchanged, so that, 2^m can be treated as a large constant and the complexity of *FIND_MaximalRandomPriorSet* algorithm is linear.

This study constructs the first basic concept about the maximal random prior set. Applying paralleled strategy to *FIND_MaximalRandomPriorSet* algorithm, along with building and comparing optimal measures of maximal random prior set are next developments of this study. This promises to increase the efficiency of the algorithm. Besides that, integrating maximal random prior set into practical applications will help verify its accuracy more clearly.

References

1. Yang, F., Guan, Y., Li, S., Du, L.: Attributes reduct and decision rules optimization based on maximal tolerance classification in incomplete information systems with fuzzy decisions. Journal of Systems Engineering and Electronics 21(6), 995–999 (2010)
2. Ravi Shankar, N., Srikanth, T., Ravi Kumar, B., Ananda Rao, G.: Genetic Algorithm for Object Oriented Reducts Using Rough Set Theory. International Journal of Algebra 4(17), 827–842 (2010)
3. Sakr, N., Alsulaiman, F.A., Valdés, J.J., Saddik, A.E., Georganas, N.D.: Feature Selection in Haptic-based Handwritten Signatures Using Rough Sets. In: IEEE International Conference on Fuzzy Systems (2010)
4. JansiRani, P.G., Bhaskaran, R.: Computation of Reducts Using Topology and Measure of Significance of Attributes. Journal of Computing 2(3), 2151–9617 (2010) ISSN 2151-9617

5. Herawan, T., Ghazali, R., Deris, M.M.: Soft Set Theoretic Approach for Dimensionality Reduction. International Journal of Database Theory and Application 3(2) (June 2010)
6. Qian, Y., Liang, J., Pedrycz, W., Dang, C.: Positive approximation: An accelerator for attribute reduction in rough set theory. Artificial Intelligence 174(9-10) (June 2010)
7. Moudani, W., Shahin, A., Chakik, F., Mora-Camino, F.: Optimistic Rough Sets Attribute Reduction using Dynamic Programming. International Journal of Computer Science & Engineering Technology 1(2) (2010)
8. Chang, F.M.: Data Attribute Reduction using Binary Conversion. WSEAS Transactions On Computers 8(7) (July 2009)
9. Wu, H.: A New Discernibility Matrix Based on Distribution Reduction. In: Proceedings of the International Symposium on Intelligent Information Systems and Applications (IISA 2009), Qingdao, P. R. China, October 28-30, pp. 390–393 (2009)
10. Lee, M.-C.: An Enterprise Financial Evaluation Model Based on Rough Set theory with Information Entropy. International Journal of Digital Content Technology and its Applications 3(1) (March 2009)
11. Parthaláin, N.M., Shen, Q., Jensen, R.: A Distance Measure Approach to Exploring the Rough Set Boundary Region for Attribute Reduction. IEEE Transactions On Knowledge And Data Engineering 22 (2009)
12. Yao, Y., Zhao, Y.: Discernibility Matrix Simplification for Constructing Attribute Reducts. Information Sciences 179(5), 867–882 (2009)
13. Liu, H., Abraham, A., Li, Y.: Nature Inspired Population-based Heuristics for Rough Set Reduction. In: Abraham, A., Falcón, R., Bello, R. (eds.) Rough Set Theory. SCI, vol. 174, pp. 261–278. Springer, Heidelberg (2008)
14. Liu, J., Mi, J.-S.: A Novel Approach to Attribute Reduction in Formal Concept Lattices. In: Wang, G., Li, T., Grzymala-Busse, J.W., Miao, D., Skowron, A., Yao, Y. (eds.) RSKT 2008. LNCS (LNAI), vol. 5009, pp. 426–433. Springer, Heidelberg (2008)
15. Parthaláin, N.M., Shen, Q., Jensen, R.: Finding Fuzzy-Rough Reducts with Fuzzy Entropy. In: IEEE International Conference on Fuzzy Systems (2008)
16. Sun, X., Tang, X., Zeng, H., Zhou, S.: A Heuristic Algorithm Based on Attribute Importance for Feature Selection. In: Wang, G., Li, T., Grzymala-Busse, J.W., Miao, D., Skowron, A., Yao, Y. (eds.) RSKT 2008. LNCS (LNAI), vol. 5009, pp. 189–196. Springer, Heidelberg (2008)
17. Yang, M., Chen, S., Yang, X.: A novel approach of rough set-based attribute reduction using fuzzy discernibility matrix. In: Proceedings of the Fourth International Conference on Fuzzy Systems and Knowledge Discovery, vol. 03 (2007)
18. Hu, Q.-H., Zhao, H., Xie, Z.-X., Yu, D.-R.: Consistency Based Attribute Reduction. In: Zhou, Z.-H., Li, H., Yang, Q. (eds.) PAKDD 2007. LNCS (LNAI), vol. 4426, pp. 96–107. Springer, Heidelberg (2007)
19. Tan, S., Cheng, X., Xu, H.: An Efficient Global Optimization Approach for Rough Set Based Dimensionality Reduction. In: ICIC International, pp. 725–736 (2007)
20. Pawlak, Z.: Rough Sets. In: The Tarragona University Seminar on Formal Languages and Rough Sets (August 2003)

SMART Logistics Chain

Arkadiusz Kawa

Poznań University of Economics,
al. Niepodleglości 10, 61-875 Poznan, Poland
arkadiusz.kawa@ue.poznan.pl

Abstract. Modern logistics companies today rely on advanced ICT solutions for information processing and sharing. Access to data and information about the demand for logistics services and supply opportunities are becoming a key competitive factor. Unfortunately, only the largest companies can afford advanced systems. Small and medium logistics companies have limited or no IT-competence. Tools are therefore needed to facilitate cooperation between smaller logistics companies, which in turn will reduce transaction costs. The paper proposes an idea of the SMART model, which is based on agent technology and cloud computing. It will allow easier collection and flow of information as well as better and cheaper access to logistics management systems.

Keywords: logistics industry, logistics services, logistics service provider, cloud computing.

1 Logistics Industry

Reliable and fast transportation, as well as efficient logistics services play a more and more important role in the activities of many companies. Logistics is not only a source of competitive advantage, but can also decide whether the firm will exist in the market at all. Therefore, a company that is focused on its core business and does not have adequate resources and experience in the logistics field is beginning to use the services of an external logistics company.

Logistics service providers are companies belonging to the so-called transport, forwarding and logistics industry. This sector covers the activities of companies of different size, multiplicity of services and global range. It includes very large but also small firms, offering a range of services - from simple transport services, through service forwarding, warehousing, palletizing, packing, packaging, to full service of supply chains. Their range of activities may comprise a region (e.g. province), country, continent or the whole world [2].

The term "transport, forwarding and logistics industry" itself points to a combination of more or less distinct activities in the past. Operators in the industry, developing new skills, fall within the competence of their more or less close competitors (carriers, freight forwarders, logistics providers). On the other hand, by working with clients, they extend their offers with additional services, sometimes taking their competitors' sales and marketing functions.

J.-S. Pan, S.-M. Chen, N.T. Nguyen (Eds.): ACIIDS 2012, Part I, LNAI 7196, pp. 432–438, 2012.

Depending on the model of business, enterprises operating on the logistics market focus on transport management (the majority of revenue comes from transport services and freight forwarding), warehouse management (inventory management services and other related services), while offering a wide range of logistics services on the basis of a contract at the same time [1, 2, 3, 5, 7].

The basis of logistics is mainly a well-developed operating system which consists of working people and infrastructure. In this system, not only parcels, documents and heavy freight are sent, but also information about them. To ensure a fast and correct flow of information between individual entities of the operating system, a logistics service provider has to use appropriate information technologies. Information and telecommunication technologies are now so closely connected with the operating system that one cannot exist without the other. Companies that are not aware of the essence of modern technology have no opportunities for further development. It is even believed now that logistics companies cannot build a competitive advantage without functioning IT systems.

Logistics companies apply information technology mainly in order to increase efficiency and automate their work. Thanks to them, numeric data presentation and fast verification and control of costs, revenues, sales and other such data clear to everyone is possible. Also, information transmission to individual business units, employees or contractors, and access to archived data is uncomplicated. Another important objective of the application of information technology by logistics service providers is to fulfill the expectations of potential and existing customers. These are the customers who care more and more about the time and safety of delivery and full information about the logistics process than about the price of the services [1, 5].

2 Computing Problems of the Logistics Industry

Unfortunately, customers using the services of various logistics service providers always have to adapt to their tools. Not only that, in the cases of one-off cooperation, they have to search for the carrier browsing their websites, trade catalogues, etc. every single time.

The logistics services industry lacks solutions that would integrate the services of different operators in one place. For example, if a customer orders goods from Taiwan to the Czech Republic, s/he has to arrange several means of transport (e.g. shipping, rail, road) and use the services of several transportation companies. A similar problem occurs in regional transport, for example, in a province or country. Although in some countries there are logistics centers (e.g. in Germany, Italy, Spain), small and medium enterprises use them to a small degree. Individual logistics centers compete with each other, ensuring comparable conditions of infrastructure. The only thing that distinguishes them is value-added services. Information about the availability of services of other businesses located near the center is an example. It includes shared information on schedules for arrivals and departures of regular and charter services, cargo tracking and tracing across organizational borders, real-time business process monitoring and exceptional situations.

Unfortunately, SME logistics companies have limited or no IT-competence and investments. They think that they should focus on business competence rather than IT [6].

The great problem of the logistics industry is a lack of formal semantics which prevents automated data integration. There are not any universal solutions which let smaller companies work together in the changing conditions and access resources, software and information provided to computers and other devices on demand.

3 SMART Model

In relation to the above questions, one should look for ways to help tackle the problem. A proposal for such a solution is the SMART model - Specialized, Market-driven, Applicable, Reactive and Technologically-oriented Logistics Chain.

The idea of this model starts from the assumption that a single company's intelligence does not necessarily imply the system's intelligence. In fact, through cooperation companies should rationalize their logistics processes, obtain cost savings and reduce empty shipments. At the moment, companies are not activating collaboration as they are traditionally managed like "family enterprises". This limits their ability to get potential opportunities offered by collaboration with other actors operating in the market.

That is why there is a need to create an electronic platform which will enable SMEs to cooperate, especially to gain access to data about logistics services and supply capacities. One of the solutions to these problems is cloud computing with web semantic services based on the Internet network (see fig. 1). It avoids capital expenditure on hardware, software, and services by paying a third-party provider only for what SMEs need to use.

Since obtaining and processing appropriate data within cloud computing is a complex and laborious process for individual actors, this research proposal suggests making use of a very promising agent technology. Software agent is a piece of software that acts for a user or another program. Agents are autonomous, thus the user can activate and disconnect them from the network, provided the agent mission is well-defined. They are capable of modifying the way in which they achieve their objectives and are, therefore, called smart agents. They are able to gather information and transform it into useful knowledge. Such agents know the data processing methods and improve them in the course of the learning process, i.e. in the course of the system operation. Smart agents are robust (they accomplish the group's objective even if some members of the group are unsuccessful), self-organized (activities are neither centrally controlled nor locally supervised), and adaptive (they respond to the dynamically changing environment). The agent community may evolve; weaker agents are eliminated and replaced by those better adapted to the market conditions.

The model proposed is based on a semantic web concept including such elements as XML, RDF, and ontologies which allow efficient automatic data collection about logistics service providers and their resources. Special logistic ontology has been created for the needs of the project (definitions of the core elements of logistics ontologies). The standard description of the content in cloud computing has been introduced in order to let software agents process data and information appropriate for their purpose and meaning. The main feature of this model is its interoperability which will enable different information systems to cooperate, safely exchange data with a predefined structure, and mutually use this data further in order to create information. What is important is that the access to such cloud computing does not require application of any specialized IT systems.

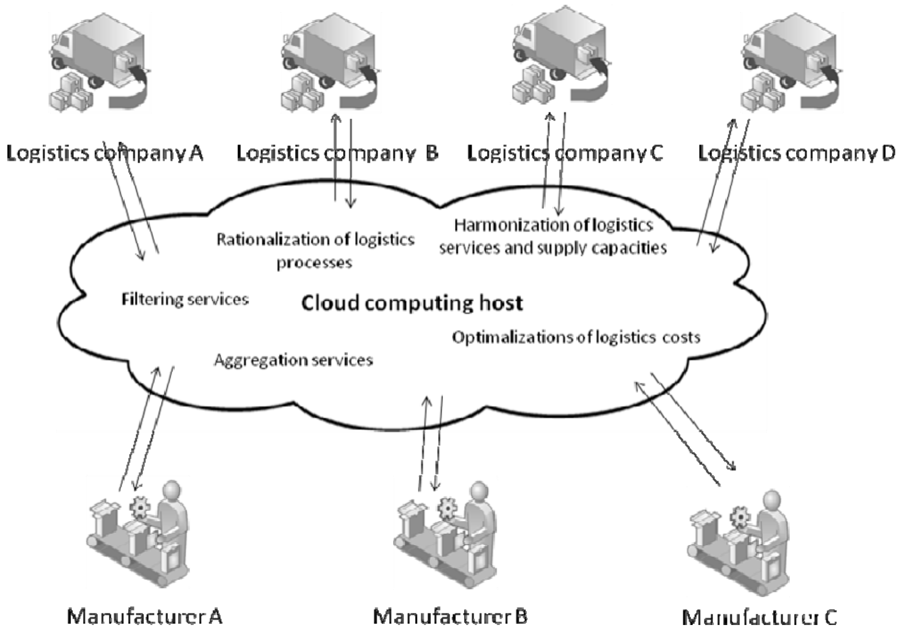


Fig. 1. Idea behind the SMART Model

Companies are provided with an electronic platform with unlimited availability and safety of its use. Relevant information is collected from various entities (e.g. providers of logistics services), filtered and aggregated on the server. Then they are made available in a suitable form to customers.

Thanks to the use of cloud computing, better and cheaper access to the systems of global logistics networks (such as DHL, UPS, FedEx) and other suppliers in this market (e.g. insurance companies, petrol stations, suppliers of car parts) will be possible.

The members of the platform will be able to optimize their cost of transport, for instance by using common transport, e.g. rail, instead of private road transport.

4 Exemplification of the SMART Model - Communication between Agents

In the SMART model, communication between the agents of individual companies within the network is very important. Agents exchange messages based on ACL (Agent Communications Language) using FIPA protocol, especially from FIPA Communicative Act Library Specification, which provides cooperation between agents systems [4]. These protocols define the syntax, semantics, and pragmatics of created messages. This solution helps to send messages between independently designed and developed agent systems [8]. Some examples of protocols of interaction between agents are presented below.

Before starting the purchase of a product, an interaction protocol (CFP - Call for Proposal) is used, in which one agent (in this case, the "Buyer" representing the buyer) submits a bid to another agent ("eMarket" for example, representing the electronic stock exchange). The buyer asks the electronic exchange environment to send proposals for sale of up to 500 tonnes of plastic for the price of less than 75 units per tonne. The presented proposal has its ID "bid09", which facilitates communication between agents by referring to it in other messages. The protocol is also a reference to the ontology which contains the specification of terms and their meanings in the field of plastics suppliers.

```
  (cfp
   :sender (agent-identifier :name Buyer)
   :receiver (set (agent-identifier :name eMarket))
   :content
     "((action (agent-identifier :name eMarket)
      (sell plastic max 500))
     (any ?x (and (= (price plastic) ?x) (< ?x 75))))"
   :ontology plastic-suppliers
   :reply-with bid09
   :language fipa-sl)
```

The recipient may accept the proposed offer or reject it, e.g. because of too low prices. If it is the first case, further communication is aimed at implementation of the proposed action; if it is the second case, the agent continues the search and sends a CFP message to other agents of the eMarket type. In the next step, the Buyer Agent, who bought the product on a commodity market, would like to have a third logistics party (3PL) to carry out the dispatch. For this purpose, it forwards the data on the numbers of the containers and shipping routes from London to Paris to the agent representing the server cloud computing (CC).

```
  (request
   :sender (agent-identifier :name Buyer)
   :receiver (set (agent-identifier :name CC))
   :content
     "((action (agent-identifier :name 3PL)
     (deliver containers000093-001956 from LON to PAR)))"
   :protocol fipa-request
   :language fipa-sl
 :reply-with order00678)
```

In the next step, in cloud computing, a suitable 3PL that can fulfill a given task is determined. The 3PL agent may agree on the task that the CC agent asks for or refuse.

```
   (agree
   :sender (agent-identifier :name 3PL)
   :receiver (set (agent-identifier :name CC))
   :content
     "((action (agent-identifier :name 3PL)
      (deliver containers000093-001956 (loc 49°31.607 'N 22°12.096'E)))
     (priority order00678 high))"
   :in-reply-to order00678
   :protocol fipa-request
 :language fipa-sl)
```

Another message that can be transferred between agents, is a request to submit additional information. The CC agent asks to send information about the services offered by 3PL.

```
query-ref
  :sender (agent-identifier :name CC)
  :receiver (set (agent-identifier :name 3PL))
  :content
    "((all ?x (available-service j ?x)))"
  :in-reply-to order00678
  :protocol fipa-query
:language fipa-sl)
```

In response to the above request, the 3PL agent informs that it offers transport, storage and packing services.

```
(inform
  :sender (agent-identifier :name 3PL)
  :receiver (set (agent-identifier :name CC))
  :content
    "((= (all ?x (available-service 3PL ?x))
      (set (forwarding service)
          (warehousing service)
          (co-packing service))))"
  :ontology logistic-services
  :in-reply-to order00678
  :protocol fipa- query-ref
:language fipa-sl)
```

Such messages sent between the agents are numerous. However, due to their similar structure, this paper has been limited to present the most important ones.

5 Conclusion and Future Work

The proposed solution gives, especially small and medium enterprises, a better possibility to automatically capture information and exchange it within a particular enterprises' network. Also, it helps to build up business relations and, finally, to better match demand and supply capacity data.

The proposed model will be tested in a simulation environment. It allows to evaluate the operating performance prior to the implementation of the system. The simulations carried out will enable companies to perform powerful what-if analyses leading them to better planning decisions and allow to compare the various operational alternatives. Moreover, it will help to understand the overall supply chain processes and characteristics by graphics/animation.

Acknowledgements. The paper was written with financial support from the Foundation for Polish Science [Fundacja na rzecz Nauki Polskiej].

References

1. Ciesielski, M.: Rynek usług logistycznych, Difin, Warsaw (2005)
2. Jeszka, A.M.: Problem redefinicji branży na przykładzie przesyłek ekspresowych. Gospodarka Materiałowa i Logistyka (7) (2003)
3. `http://www.businessmonitor.com`
4. `http://www.fipa.org/specs/fipa00037/SC00037J.html`
5. `http://www.kep.pl`
6. `http://research.microsoft.com/en-us/people/sriram/rehof-cloud-mysore.pdf`
7. `http://www.transportintelligence.com`
8. Kawa, A., Pawlewski, P., Golinska, P., Hajdul, M.: Cooperative Purchasing of Logistics Services among Manufacturing Companies Based on Semantic Web and Multi-agents System. In: Demazeau, Y., et al. (eds.) Trends in PAAMS. AISC, vol. 71, pp. 249–256. Springer, Heidelberg (2010)

The Analysis of the Effectiveness of Computer Assistance in the Transformation of Explicit and Tactic Knowledge in the Course of Supplier Selection Process

Karolina Werner, Lukasz Hadas, and Pawel Pawlewski

Poznan University of Technology, Strzelecka 11, 60965 Poznan, Poland,
{karolina.werner,lukas.hadas,pawel.pawlewski}@put.poznan.pl

Abstract. The article presents investigations upon the effectiveness of computer assistance in the transformation of explicit and tactic knowledge in the course of supplier selection process in an enterprise. A computer system is oriented towards achieving greater process efficiency than in a traditional procedure based upon expert knowledge of employees. Authors of the article have modeled a multi-variant process of supplier selection providing two options, one with computer assistance and the other without this assistance. The obtained results are used to identify the restrictions and formulate the conclusions as for the requirements that should be met by the systems assisting the decision-making in this area.

Keywords: knowledge transformation, assisting the expert decision-making.

1 Introduction

The process of transforming tactic knowledge into explicit knowledge is of a complex nature. The difficulty of studies upon the knowledge flow results from the necessity of constructing structural and functional models with known parameters obtained from precise procedures of model-identification, as well as the way of defining the parameters and the need for the measurements and evaluations. Such models are the basis of simulation, i.e. the virtual investigating the transforming the reality [2].

The article presents results of an experiment concerning transforming tactic knowledge into explicit knowledge. The object of the process, i.e. the essence of the flow, is in fact difficult to grasp and define. The presented investigations are based upon a discrete model of simulation, ascribing a process to an object identified as a so-called knowledge unit. This is an abstract concept, identified by means of defining an evaluation set. The evaluation set contains criteria that make it possible to identify the object of a process, i.e. knowledge units, by listing the consecutive states after the transformation. This conception enables using the discrete model of simulation to perform the experiments.

As a result of the investigations, it is possible to analyze, both from the quantitative and temporal perspective, the efficiency of knowledge transformation process for the

J.-S. Pan, S.-M. Chen, N.T. Nguyen (Eds.): ACIIDS 2012, Part I, LNAI 7196, pp. 439–448, 2012.

particular paths of implementing the analyzed function of supplier selection, so as to meet the requirements generated by a production department. The simulation included two variants of implementing the function: in a traditional way and with computer assistance in the form of electronic purchase platform.

2 Knowledge Transformation

2.1 The Essence of Knowledge Transformation

The role of knowledge in an economic activity, and especially in management, has been discussed for decades [7]. Hoverer, knowledge began to be considered [4] as a subject of management in organizations no sooner than in the last ten years of the 20th century. Since that time both science and business practice have provided a number of instructions so as to obtain, master, develop and use this knowledge.

Knowledge is not the same as information, although these two notions are commonly treated as synonyms. It can be assumed that knowledge is information that has been understood, enriched with opinion and used in operation. Knowledge is information in context along with understating how to use it [7]. The understanding emerges from the resources of knowledge possessed by people or/and institutions [7].

In management, knowledge is usually treated as [8]:

- connecting information with its understanding,
- the effect of intellectual processing of information and experiences and learning,
- the total of a person's knowledge,
- the refection of reality in a human mind,
- confirmed belief.
- Knowledge treated as a resource to be used in business is characterized by the following [8]:
- knowledge can be created by various means,
- it cannot be easily grasped or fully used but, contrary to material resources, it can be used by various people in various places,
- it is relative and ambiguous and therefore it can interpreted differently be various people,
- it is dynamic – grasping a part of it can lead to its massive increase while it is being used,
- it can quickly become obsolete
- it can decrease the uncertainty level in risky ventures,
- owing to the codification process it is structured in technologies, procedures, documentation, employees competence and databases,
- it can materialize, which means that it is manifested in products and services and therefore it is imitable (with various levels if difficulty though),
- knowledge itself can be a product.

Due to all of the above, knowledge is a resource that cannot be easily represented in a simulation process and it is difficult to assist its transformation with use of IT tools.

2.2 Explicit and Tactic Knowledge in the Transformation Process

The first researcher who distinguished between explicit and tactic knowledge was M.Polanyi, who stated 'We can know more than we can tell' [11].

Tactic knowledge is individual, specific and contextual, difficult to formalize and communicate and requires specific learning skills [1].

S. Chełpa lists the following features of tactic knowledge [3]:

- individualization and personification, which means that knowledge results mainly from experience, and that is why it is greatly subjective and relative,
- idiosyncrasy, which means contextual sensitivity and a tendency to be revealed in certain external conditions,
- pre-theoreticalness and non-intellectualism, which means that its arrangement and occurring modifications do not result from an intentionally initiated process of thinking,
- internalization and automation, which means that knowledge is characterized by deep interiorization, and revealing it is proceeded automatically (instinctively) beyond the control of consciousness, although the tasks are target-oriented,
- difficulty in localizing: it eludes the attempts to codify it conceptually and transferring it verbally to other people.
- The classification that makes a distinction between explicit and tactic knowledge was expanded by Japanese researchers I. Nonaka, R. Toyama and N. Konno, who distinguished the following [11]:
- experimental knowledge assets, which result from common experience and education of individual people and consist of individual skills and know how as well as energy, passion, mutual trust, help and sense of security,
- routine knowledge assets, i.e. knowledge transformed into routine by means of practical actions; it consists of abilities necessary to perform certain activities as well as the routine of organizational behavior and organizational culture,
- conceptual knowledge assets, which mean formal and specified knowledge expressed in the form of images, pictures, symbols and language (e.g. specific projects, patterns, models),
- systemic knowledge assets, i.e. knowledge formalized in the form of documents, instructions, databases, patents etc.

These four groups are considered as knowledge assets, which are created, developed and one by one transformed into the spiral SECI process, that consists of socialization, externalization, combination and internalization [7]. According to m. Warner and M. Witzel the transformation of knowledge is a process which is the essence of management: knowledge resources, i.e. gathered data, analyses and inspirations, gain ultimate value.

3 The Course of Simulations – Creating Process

3.1 Methodology of the Investigations

As far as knowledge transformation is concerned, simulation is a form of experimenting on a computer model (simulation model) that answers the question how the analyzed processes related to knowledge transformation will behave in the given situation spectrum. It makes it easier to understand the functioning of the modeled processes. Considering the process model as an actual duplicate makes it possible to implement conclusions from the investigations based on a computer model into processes in accordance with the principles of model construction. Each model should be treated individually because usually it cannot be used to analyze another decision-making problem.

The course of modeling and simulation experiment can be divided into three stages [10]:

Stage A – constructing a conceptual model of simulation. At this stage, a problem is formulated and a research plan is prepared. It is necessary to determine the objectives of the project and the reasons for using simulation as well as to define attributes to be evaluated and quality measures of the solution. Based on that, the range of the project and the level of accuracy should be determined. In the conceptual model of simulation, the key components of a problem and their relations are identified, and the transformation of uncontrolled and controlled inputs into the set data output are described.

Stage B – constructing the computer simulation model. At this stage, the computer model is built step by step. First, a simple, correctly operating model is built and checked. Next, it is extended gradually. After each step, the operation of the model is examined. This method of structuring is recommended by manufacturers of simulation software [5].

Stage C – carrying out simulation experiments. The next stage after constructing a properly operating computer model, is its validation and designing an experiment. In the experiment project the initial conditions must be determined and all decision variables, that may be of interest to the user and that will be analyzed in separate courses, must be predicted. In order to reduce the number of experiment repetitions and simulation courses, without deteriorating the quality of the obtained results, the methods of planning an experiment can be applied. The fact that, during the implementation of the modeling and simulation, all the current results are being analyzed should also be taken into account; depending on the quality of these analyses, recurrences of the taken activities may occur.

Validation of the simulation model is one of the crucial stages of implementing the simulation experiment, since it makes it possible to determine the reliability level of the results. Validation should be distinguished from verification, which is checking the correctness of formal model transformation into the form of a computer program.

3.2 Conceptual Model of Simulation

The conceptual model of simulation was constructed on the basis of the methodology of building process flow model [8], which uses elements of IDEF0 methodology. However, the main component of the prepared conceptual model is the so-called concept of a process object [9]. A process object (i.e. aim of a process) can be a material product, a document (information product), decision (information product). The suggested solution ascribes a process to an object. Therefore, the process can be unambiguously recognized and the state of the process can be clearly determined so that one can precisely identify the current stage of the process. It is very important to take into account the time here, i.e. to consider the process in time. In order to identify the current stage of a process, it is necessary to identify the place where, in the given time, the so-called process object is present. In order to identify the process of knowledge transformation, the so-called knowledge unit was introduced as a process object. Further in the article, it is identified as KU. Knowledge unit is an abstract concept which can b identified by defining the so-called evaluation set.

$KU \in \{KU1, KU2, KU3, \ldots., KUn\}$

where: $n \in N$.

The developed conceptual model consists of:

- a process map
- a process chart based on IDEF0 methodology
- a table describing the evaluation set KU
- components enabling and showing development directions of the model.

A process map is shown in Figure 1. The map illustrates activities in the process and shows the flow of KU through the process.

The model of mass support was selected for the implementation of a given activity. In the mass support system a demand occurs (in this case, it is KU) along with a demand for support. The system of massive support reacts by meeting the requirement if possible. Otherwise it retains the demand until the right moment and, consequently, queues appear in front of operating stations. The queues are formed by various tasks. At a given moment, they are represented by one specific operation which is to be performed at the given station. The queue is a set of specific individual operations that belong to various tasks waiting in front of an operating station. The characteristics that describes the mass support models refers to the following:

- stream of demands
- service process
- the way a queue is serviced.

The stream of demands is described with use of a time interval between consecutive demands. The interval may be constant. In case, the demands are random, the interval is also a random variable and its probability function should be determined. The cumulative probability functions, that determine the probability that a time interval is larger than a certain time value, are commonly used. In the analyzed case, the stream of demands is passive, which should be interpreted as an infinite queue of demands which are selected for the first operation as soon as possible.

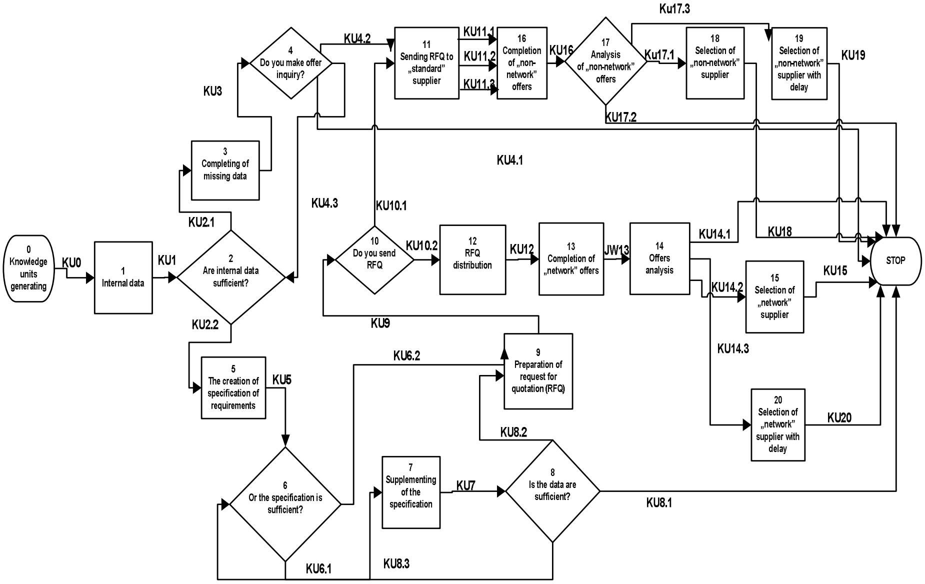

0
Knowledge units generating
KU0
1
Internal data
KU1
2
Are internal data sufficient?
KU2.1
3
Completing of missing data
KU3
4
Do you make offer inquiry?
KU4.2
KU4.3
KU4.1
KU2.2
5
The creation of specification of requirements
KU5
6
Or the specification is sufficient?
KU6.1
KU6.2
7
Supplementing of the specification
KU7
8
Is the data are sufficient?
KU8.1
KU8.2
KU8.3
9
Preparation of request for quotation (RFQ)
KU9
10
Do you send RFQ
KU10.1
KU10.2
11
Sending RFQ to „standard" supplier
KU11.1
KU11.2
KU11.3
12
RFQ distribution
KU12
13
Completion of „network" offers
JW13
14
Offers analysis
KU14.1
KU14.2
KU14.3
15
Selection of „network" supplier
KU15
16
Completion of „non-network" offers
KU16
17
Analysis of „non-network" offers
Ku17.1
KU17.2
Ku17.3
18
Selection of „non-network" supplier
KU18
19
Selection of „non-network" supplier with delay
KU19
20
Selection of „network" supplier with delay
KU20
STOP

Fig. 1. Map of KU transformation process in the process of supplier selection

Service process describes how a demand is handled. The process is determined by two factors: service time and multiplicity of the support system. Service time is the time required to support one demand. Multiplicity of the support system is the number of demands which can be handled simultaneously.

The third aspect is the system accessibility. In the analyzed case, each activity (service process) is related to a description of service time. In most cases, it is presented as specific probability distribution. The multiplicity of the support system in all cases amounts to one. Therefore, multitasking is not being considered. The system is always accessible.

4 Simulation Experiments

The simulation referred to the process of supplier selection on the basis of available knowledge about offered products. The simulation was restricted to 1600 man-hours, so as to reflect effective annual standard hours of a single worker. The simulation was an experiment and that is why, the analyses were restricted to studying the process efficiency. In order to thoroughly analyze the whole process, it would be necessary to imitate the functional division of labor in the investigated company.

Knowledge units (KN) were generated only at the beginning of the process. While analyzing the process time, in order to, e.g. obtain the missing data, it turned out that it is relatively long and it may become a bottleneck. In reality, there may be several operating stations that implement the same tasks. Then, the multi-streaming character of the process will restrain it from becoming a bottleneck. The FIFO rule (first in, first out) is kept for input and output of the process, because the analysis of the plans implementation was not the objective of the investigation. Before each input and output, a queue of KN was formed and then an automatic machine took the first KN, that had been waiting for the longest time.

The experiments were carried out for the version with a traditional procedure and for the one that is computer-assisted. For comparison purposes, for both versions the same streams of random numbers were used. They were changed when the experiment number was changed.

5 Analysis of Simulation Results (Fragments)

After the computer model of simulation was completed and validated, a series of experiments were carried out in order to confirm the results of earlier studies and draw ultimate conclusions. After the operation of the simulation model was positively verified in this aspect, generating of the courses began so as infer about knowledge transformation in the process of supplier selection. For the detailed inference process, 6 simulation courses were chosen both in the computer-assisted variant and in the traditional mode of generating a request for proposal**,** and selecting the final supplier.

A comprehensive view on the residence time in buffers and the lead time of operations in the “system with ON” and “without ON” for all 50 knowledge units shows (Figure 2) that, in most cases of units compared in pairs, time and labor consumption is smaller for the variant with computer-assistance.

Nevertheless, the obtained effect of shortening the time of implementing the function of supplier selection based on knowledge transformation is weaker than expected. IT system (represented by Open Nexus) assists the following:

- gaining information from databases
- preparing a request for proposal (according to a template)
- internal distribution of requests (between employees in a company)
- external distribution of requests to potential customers
- collection of requests (automation of the process)
- automatic analysis of the requests and selection of the best offer (based on criteria determined the moment the request for proposal was made).

Therefore, greater benefits, such as decrease in the time-consumption of lead time, were expected. A more precise analysis, however, revealed the reasons for such results.

In the chain of operations performed in the analyzed process, the effect of bottleneck appeared. That significantly reduced the flow of knowledge units. What is more the bottleneck effect occurred in the initial phase of the process, and therefore it was impossible to “fill” the whole system with knowledge units at the proper level. This would probably reveal larger differences between the considered variants of implementing the function. Because of the attempt to imitate the process that exists in reality, the researchers did not try to balance the flow capacity, since it would not comply with the validated model.

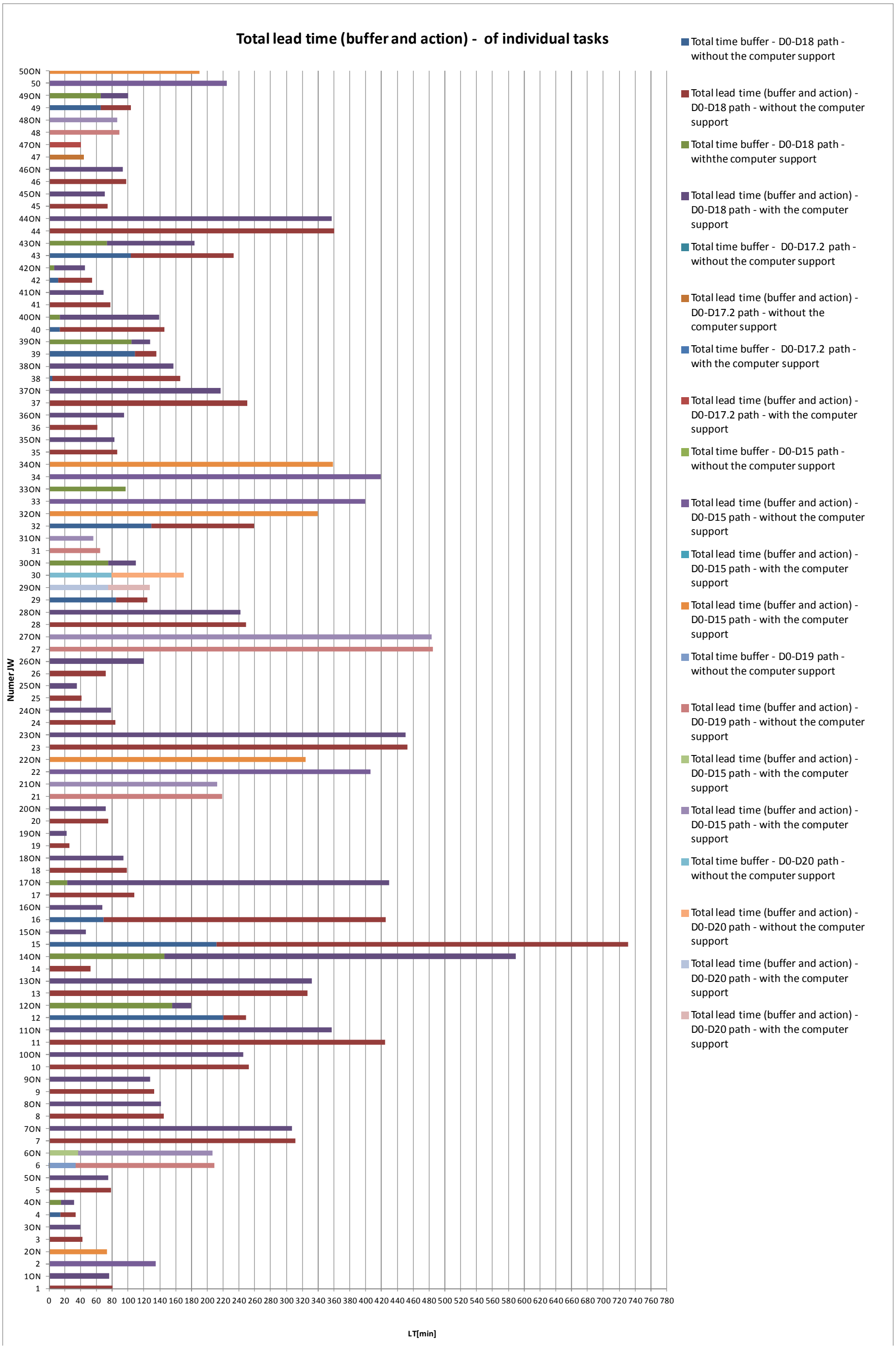


Fig. 2. Comprehensive analysis of residence time In the buffers, lead time of an operations and time of residence In the system with ON and without ON for all knowledge unit.

6 Conclusions

The imitation of the transformation process was based on specific organizational conditions of an enterprise with high level of ready product customization. In such conditions, owing to the fact that demand for various products is greatly changeable in time, there is a great necessity for the transfer of knowledge concerning purchase needs, which are often non-standard and unique. Due to the complexity of quality aspects of the transformation process, the validation of the imitated process was restrained only to the investigations upon the qualitative aspects of implementing the supply function.

The concept of knowledge unit was introduced as information carrier, that is subjected to transformation. This is a unit which simulates the flow of knowledge in the chain of operation sequence and can be addressed to and processed by numerous subjects. In the process of analyzing the obtained results, time-effectiveness (i.e. time and labor consumption) of the function implementation was evaluated for the system with and without computer-assistance.

The results of analyses confirm the potential of IT systems in assisting the process of knowledge transformation, distribution and storage. Nevertheless, a greater effort must be put into supporting knowledge transformation with use of expert competence, since the process is crucial for constructing the proper specification of the request for proposal.

It should be noted that the phenomenon of bottlenecks (occurring in the phase of tactic knowledge transformation) was identified as a resource that is timing the processing of knowledge units at the consecutive stages of the sequence. Therefore, the phenomenon must be taken into consideration while designing IT solutions [11,12]. Ignoring this fact will result in lacking the possibility of achieving greater efficiency of the system in terms of knowledge processing (implementing functions for various needs).

The problem mentioned above leads to another issue. In the conditions of a real company, the implementation of the sequence of particular tasks is changeable. The input sequence of operations in the system may change, based on the corrections that are made (controlling the process) due to occurring delays, changes in implementation priorities or assigning different resources to the implementation of particular tasks. In this way congestions (queues of operations to be performed) can be removed and possible delays can be avoided. Thus, the designers of IT system assisting the process of knowledge transformation ought to develop a mechanism of:

- dynamic allocation of resources for implementing particular tasks
- monitoring and evaluating the progress level of implementing particular tasks taking into account the priority level.

The implementation of the above-mentioned suggestions will make it possible to increase the efficiency of computer programs assisting the process of knowledge transformation in business organizations.

References

[1] Boiral, O.: Tacit Knowledge and Environmental Management. Long Range Planning 35, 296 (2002)

[2] Cempel, C.Z.: Nowoczesne Zagadnienia Metodologii i Filozofii Badań, ITE Radom, Poznań, p. 62 (2005)

[3] Fertsch, M., Pawlewski, P.: Comparison of process simulation software technics, Modelling of modern enterprises logistics. In: Fertsch, M., Grzybowska, K., Stachowiak, A. (eds.), Monograph, p. 29, 39. Publishing House of Poznań University of Technology, Poznań (2009)

[4] Grudzewski, W.M., Hejduk, I.K.: Zarządzanie wiedzą w przedsiębiorstwach, Difin, Warszawa, p. 22 (2004)

[5] Law, A.M., Kelton, W.D.: Simulation Modeling and Analysis, p. 13. McGraw-Hill, New York (2000)

[6] Mikuła, B.: Zarządzanie wiedzą w organizacji, Podstawy zarządzania przedsiębiorstwami w gospodarce opartej na wiedzy, Difin, Warszawa, p. 113 (2007)

[7] Mikuła, B., Pietruszka-Ortyl, A., Potocki, A.: Zarządzanie przedsiębiorstwem XXI wieku, Difin, Warszawa, p. 69–72, 129 (2002)

[8] Pawlewski, P.: Budowa modelu przepływu procesu, Instytut Inżynierii Zarządzania. Politechnika Poznańska, pp. 1–10 (2006), `http://www.wbc.poznan.pl`

[9] Pawlewski, P.: Metodyka Modelowania Dynamicznych Zmian Struktury Zasobowej Procesu Produkcyjnego w Przemyśle Budowy Maszyn. Rozprawa habilitacyjna – maszynopis, Poznań, p. 97 (2011)

[10] Rojek, M.: Wspomaganie procesów podejmowania decyzji i starowania w systemach o różnej skali złożoności z udziałem metod sztucznej inteligencji, Wyd. Uniwersytetu im. Kazimierza Wielkiego, Bydgoszcz, p.4 (2010)

[11] Tzung-Pei, H., Cheng-Hsi, W.: An Improved Weighted Clustering Algorithm for Determination of Application Nodes in Heterogeneous Sensor Networks. Journal of Information Hiding and Multimedia Signal Processing 2(2), 173–184 (2011)

[12] Wang, T., Liou, M., Hung, H.: Application of Grey Theory on Forecasting the Exchange Rate between TWD and USD. In: International Conference on Business and Information, Academy of Taiwan Information System Research and Hong Kong Baptist University, Hong Kong, July 14-15 (2005)

Virtual Logistics Clusters – IT Support for Integration

Paulina Golinska[1] and Marcin Hajdul[2]

[1] Poznan University of Technology, Strzelecka 11, 60965 Poznan, Poland
paulina.golinska@put.poznan.pl
[2] Institute of Logistics and Warehousing, Estkowskiego 6, 61755 Poznan, Poland
marcin.hajdul@ilim.poznan.pl

Abstract. Companies are facing problems regarding reduction of their logistics costs. Good organization of transport processes can bring a lot of profits both economical and environmental. Companies participating in supply process more and more often prefer to create temporary relations and form virtual cooperation networks than to keep traditional long-term contracts. The aim of the paper is to present the tool that supports the coordination of transport in virtual logistics clusters. The main problems and requirements are identified. Moreover authors present the results of questionnaire conducted among transport users and transport providers regarding the communications standards.

Keywords: logistics, transport processes, cooperation, communication standards.

1 Introduction - Virtual Logistics Clusters

Small and medium enterprises usually have less money for infrastructure expenses. The web-based services allow overcoming this limitation. The common access to broadband Internet usually is the only required infrastructure to use them. Moreover web-based solutions allow integration among group of companies in order to form temporary cooperation network.

Traditionally clusters are understood as geographic concentrations of interconnected companies, specialized suppliers, service providers, and associated institutions in a particular field that are present in region [1]. Virtual logistics cluster can be defined as a network of individual companies that are organized around the particular interconnected logistics process which is performed in a geographic region. Transport processes are good example of such cooperation, because they are usually dedicated to transfer goods to particular geographical location (region).

The aim of the virtual logistics cluster is to achieve the effect of the economy of scale and reduction of cost of the delivery process. In case of transport processes there are some additional benefits, like:

- better utilization of vehicles: increased load factor and reduction of empty routes,
- reduction of greenhouse gasses due to optimization of route planning.

The virtual logistics clusters are described by a group of common characteristics:

J.-S. Pan, S.-M. Chen, N.T. Nguyen (Eds.): ACIIDS 2012, Part I, LNAI 7196, pp. 449–458, 2012.

- technology- application of modern IT tools and developed telecommunications infrastructure provide exchange of data necessary for the effective organization of logistics processes,
- opportunistic approach to network configuration – any time companies evaluate the benefits of participation in the virtual organization and compare them with benefits of individual performance,
- temporary character of relations- usually there is no long-term contracts and companies are able the reconfigure the network by each contract.

The temporary nature of the cooperation requires common standards for communication which are easily applicable at low costs. Some standards already have been elaborated but still a shift in interoperability is needed. One example of such industrial initiative is the development of a new Logistics Interoperability Model (LIM) by GS1. This model is elaborated and applied mainly by big companies. Small and medium companies still search for the other solutions that might be suitable for them.

In the next sections authors describe the requirements for the integration tool, as well as current situation regarding the implementation of common communication standards among companies. The transport process is analyzed as an example.

2 Integration of Transport Processes and Challenges for Communication within Cluster

The main issue that appear by coordination of process in the virtual cooperation network is independency of technology in order to support the possibility of temporary relations and reconfiguration.

In case of the transport process the requirements for the Common Framework have been identified [2], as:

- support multimodality (co-modality),
- be stable and easy to refine and expand,
- be future-oriented (independent of current solutions),
- provide a total picture (supporting transparency, management, and security),
- facilitate hiding of complexity (abstraction, simplification),
- focus on interoperability (not on inner parts of systems),
- independent of technology,
- facilitating interaction with existing standards (to help protect investments already made in legacy and other systems).

In case of transport very often is used electronic data interchange (EDI) including special message formats elaborated for transport of goods like for example:

- IFTMIN (instruction transport message) – an EDIFACT message from the client regarding forwarding/transport services for a consignment under conditions agreed. The message is from shipper to carrier or forwarder containing the final details of the consignment for which services are provided.
- IFTSTA (International multimodal status report) - an EDIFACT message to report the transport status and/or a change in the transport status (i.e. event)

between involved companies. It reflects information on the status of the physical movement of consignments, goods at any point.

2.1 Reference Model of Cooperation within Virtual Logistics Cluster

The paper presents the creation of a comprehensive tool supporting the consolidation of transport processes for group of independent companies. The aim of cooperation is optimizing transport costs resulting from economies of scale achieved through sustainable exploitation of resources in the group. The benefit of co-operation is also reducing the cost of staff associated with the organization of transport by transferring some responsibilities to a coordinator of transport.

The main task for the reference model is to stimulate cooperation between group of manufacturing (users) and transport enterprises (providers). The concept is based on idea of sustainable development. Figure 1 presents reference model.

Cooperating entities shall exchange information electronically via dedicated electronic platforms. In the reference model three groups of players are defined (see fig. 1):

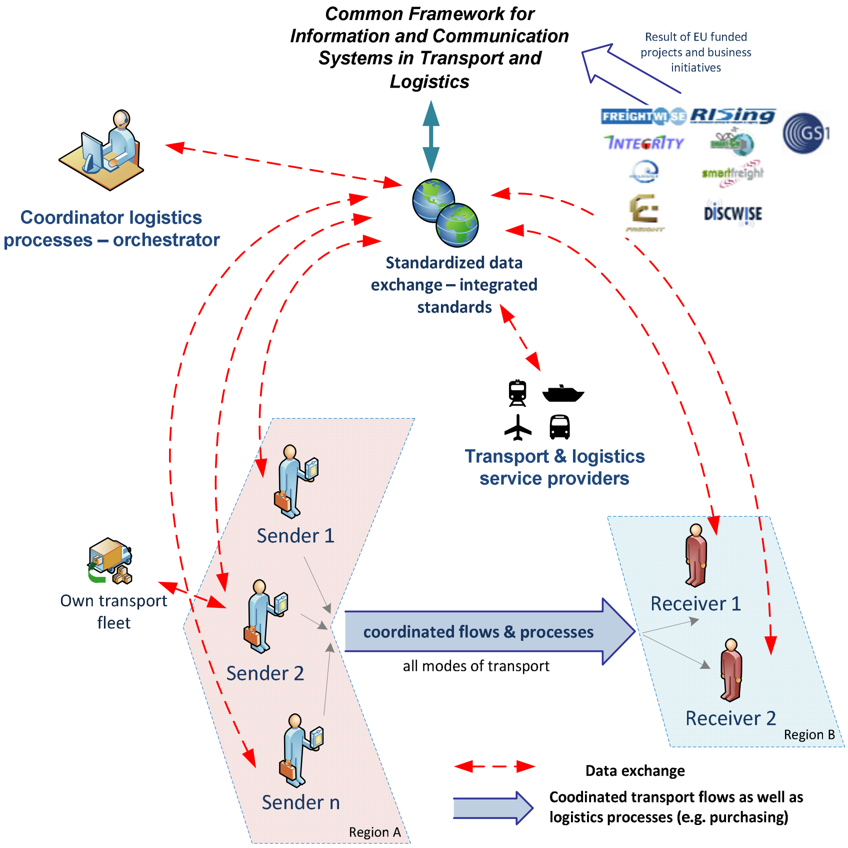

Common Framework for Information and Communication Systems in Transport and Logistics
Result of EU funded projects and business initiatives
FREIGHTWISE
RISING
INTEGRITY
GS1
smartfreight
DISCWISE
Coordinator logistics processes – orchestrator
Standardized data exchange – integrated standards
Transport & logistics service providers
Own transport fleet
Sender 1
Sender 2
Sender n
Region A
coordinated flows & processes
all modes of transport
Receiver 1
Receiver 2
Region B
Data exchange
Coodinated transport flows as well as logistics processes (e.g. purchasing)

Fig. 1. Reference model for transport process performed by virtual cluster **[3]**

- Users of transport services – companies that are engaged in the production and/or selling of the products. Logistics is not their main source of business. These companies may have their own means of transport, logistics infrastructure or they cooperate with the providers of logistics services. They issue the demand for transportation. They order the execution of logistics operations in most of the cases by issuing order to providers of services.
- Provider of transport services – companies whose core business is the provision of logistics services. Their task is the coordination of orders issued by logistics users. In the case where one of the cooperating firms has its own facilities, e.g. means of transport and is able to provide transport services for the other transport users, then it performs also the role of transport services providers.
- Coordinator – represents the users and deals with the coordination of logistics processes (e.g. analysis of the possible aggregation of the transport orders issued by the different users, price negotiations, and the choice of modes of transport).

Developed reference model takes into consideration [4]:

- assumptions for the European transport policy development,
- co-modality of transport processes,
- one common standard in transport & logistics data exchange which has been developing by several EU research projects such as FREIGHTWISE, e-Freight, INTEGRITY, Smart-CM, SMARTFREIGHT, DiSCwise, COMCIS, iCargo,
- correlation between logistics system of the enterprise and the regional/national/continental transport systems,
- traceability issue, like GS1 [5],
- possibility of using particular computing tools and information exchange techniques.

Within the framework of the model the following instruments stimulating cooperation between enterprises need to be addressed during implementation phase:

- methodology enabling estimation of potential savings and profits that derive from cooperation among enterprises in logistics processes organization,
- legal framework for cooperation between shippers, logistics service providers and co-ordinator (orchestrator),
- methodology that assure common planning of logistics processes in a group of enterprises taking into consideration relation of a trade-off between transport, inventory management as well as warehousing processes,
- methodology that assures common planning of transport processes in a group of enterprises in compliance with a trade-off relation between micro scale (enterprise) and macro scale (region),
- exploitation of existing e-platforms (T-Scale, Logit 4SEE) that support interoperability and harmonizing logistics processes in group of enterprises and as a result allows joint organization of haulages. Furthermore the e-platforms are to support standardized data exchange process thereby enabling cooperation between users and transport service provider for all modes.

2.2 Survey Results

The pilot survey was conducted among group of 40 companies. The companies represent both transport users and transport providers including logistics operators and forwarders. First group of questions (see fig.2) concerns the present situation regarding the application of IT tools as well as willingness to implement new platform.

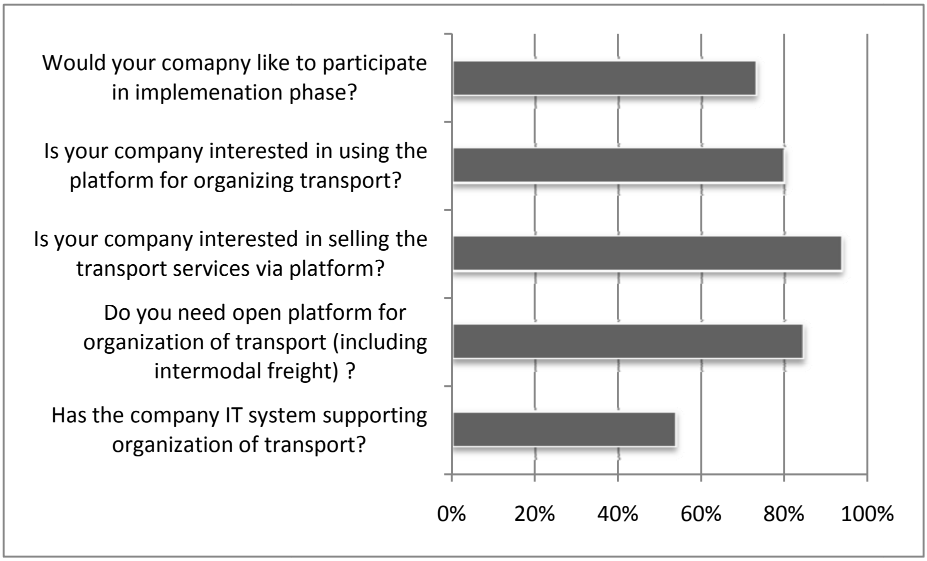


Fig. 2. Positive answers regarding the IT support of transport processes

In the analyzed group 54% of companies have IT systems which support they everyday operations by planning and organization of transport processes, but still the companies see the need to use the integration platform for cooperation among transport providers and transport users (80% positive answers). Over 90% of transport providers (including logistics operators) were interested in selling they services via integration platform. In the open question regarding factors influencing the willingness to use the platform the most popular answers were:

- cost reduction,
- better resource utilization,
- flexibility,
- reduction of time needed for communication,
- less errors occurring due to lack of information or inappropriate information exchange.

Additional interviews were conducted with 18 transport users. The answers are presented in figure 3.

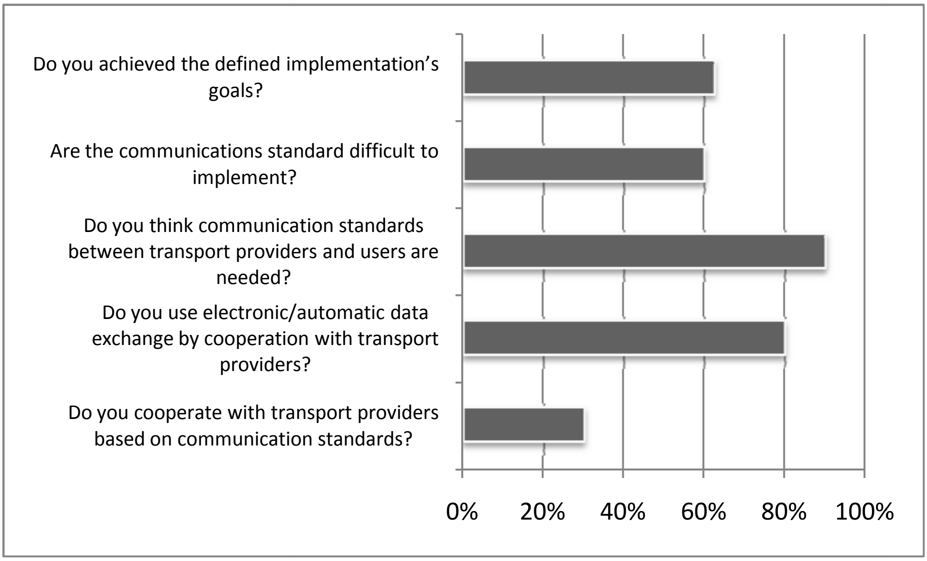


Fig. 3. Positive answers regarding the application of common communication standards - users

At present the analysed companies don't use very often common communications standards (30% of positive answers). Most of the companies use electronic data exchange. In the open question the respondents have identified the main barriers which appear by implementation of communications standards:

- different levels of infrastructure development between cooperating companies,
- staff problems (lack of technical competences),
- different organizational culture of companies.

Most of transport users have declared that common communication standards are needed. The main advantages of implementation of communications standards defined by users are:

- quality improvement of the transport services,
- costs optimization,
- reduction of delivery times,
- simpler process flow,
- reduction of time needed for communication.

Additional interviews were conducted with 22 transport providers. The answers are presented in figure 4.

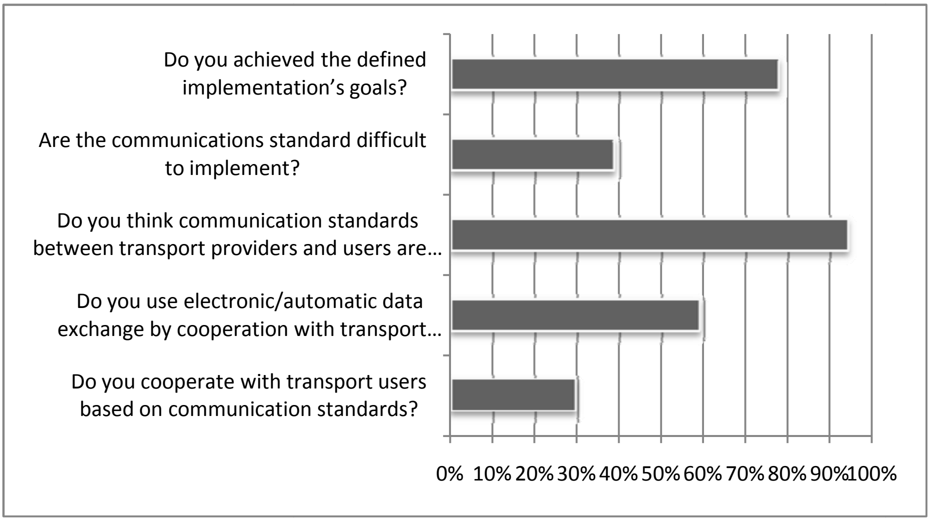


Fig. 4. Positive answers regarding the application of common communication standards – transport providers

At present the analysed companies don't use very often common communications standards (30% of positive answers). Most of the analysed transport providers use electronic data exchange (c.a.60%). In the open question the respondents have identified the main barriers which appear by implementation of communications standards. From transport providers' point of view the typical barrier is:

- high cost of implementation,
- long time needed for implementation,
- human factor (difficulties with the staff training),
- technical difficulties (system errors) in the initial stage of implementation.

Most of transport providers (94%) have declared that common communication standards are needed. The main advantages of implementation of communications standards defined by providers are:

- reduction of errors by order fulfilment,
- cost reduction,
- reduction of time needed for communication,
- transparency of information flow,
- access to information in real-time.

In the next section authors describe the integration tool called T-Scale, which helps the companies to form virtual transport clusters.

3 Integration Tool

T-Scale is an IT tool which is enabling the exchange of information and plan missions in real-time between actors involved in the carriage (transport user, the provider of transport services, coordinator). Cooperation of independent companies is also a way to increase the availability of the loading capacities in trucks. Another advantage is the reduction of traffic, thus reducing the negative impact of road transport on the environment.

Unlike the existing transport's e-market the tool gives the opportunity for coordination and consolidation of orders, thus optimizing transport costs arising from the achieved economies of scale. It increases trust and security resulting from cooperation in a closed group of companies and monitoring of cooperation by the independent entity (virtual cluster). Figure 5 presents the simplified information flow within the virtual cluster.

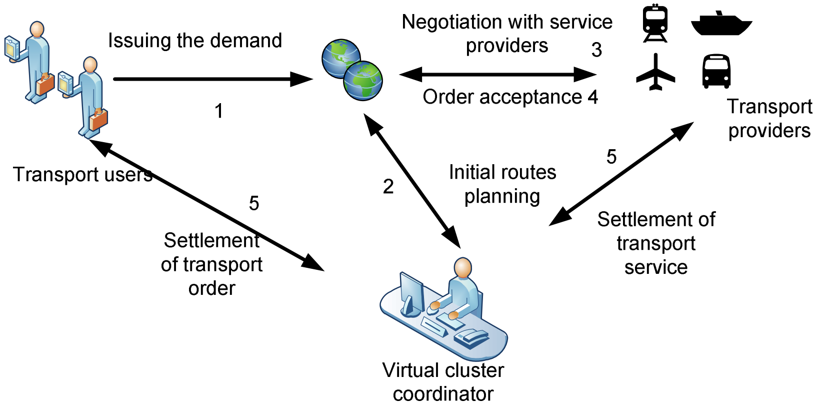

Issuing the demand
1
Negotiation with service providers
3
Order acceptance 4
Transport providers
Transport users
5
Settlement of transport order
2
Initial routes planning
5
Settlement of transport service
Virtual cluster coordinator

Fig. 5. Simplified information flow

The aim of the tool is to manage the transport fleet within the cluster in order to eliminate empty routes on the way back and/or consolidate transport demand to joint location.

The initial test of proposed tool was carried out in pharmacy and FMCG (fast moving consumer goods) sector within DiSCwise project, founded by DG Enterprise, European Commission in 2010-2011. The project aims to Develop, Demonstrate and Deploy a Reference Architecture for Interoperability in the Transport and Logistics Sector in an effort to achieve:

- integration of small and medium sized transport service providers into efficient door-to-door supply chains at cost affordable to them,
- facilitating a more sustainable transport for users to select environment-friendly alternatives.

The initial test was performed in virtual pharmacy cluster in Poland. The aim was to support cooperation between suppliers, wholesalers, logistics service providers in order to reduce transport costs through reduction of empty runs and aggregation of deliveries of independent companies. One of the major constraints was not to decrease customer service level.

Reduction of transport costs by increasing the level of exploitation of the load space, while maintaining the required frequency of supply. Combining volumes of deliveries carried out by only two companies in exactly the same days, for exactly the same location has enabled cost savings achieved at the level of the ca. 18% of the costs incurred so far. Performed analysis showed that the great potential for reducing costs by means of a common organisation of the supply is in:

- aggregating volumes of deliveries regionally,
- defining common customers' locations by region, connecting the supply to the common location in the few days period-negotiating conditions for the supply of customers in joint locations.

First tests were carried out on a small sample representing group of four distributors. Before implementation of new organisational solution, companies did not cooperate at any level, but had deliveries to the same clients or clients located very close to each other. In analysed case illustrating average day, each of the chemists ordered selected amount of products, which were delivered in special boxes (ca. 30), dedicated and standardized for pharmacy sector. Due to application of IT support for forming virtual transport cluster the number of deliveries was limited at the same time the required customer service level was secured. It is worth paying attention to is that number of trucks is reduced what decreases road traffic congestion. Therefore, the solution eliminates disadvantages of the traditional method of transport process organization.

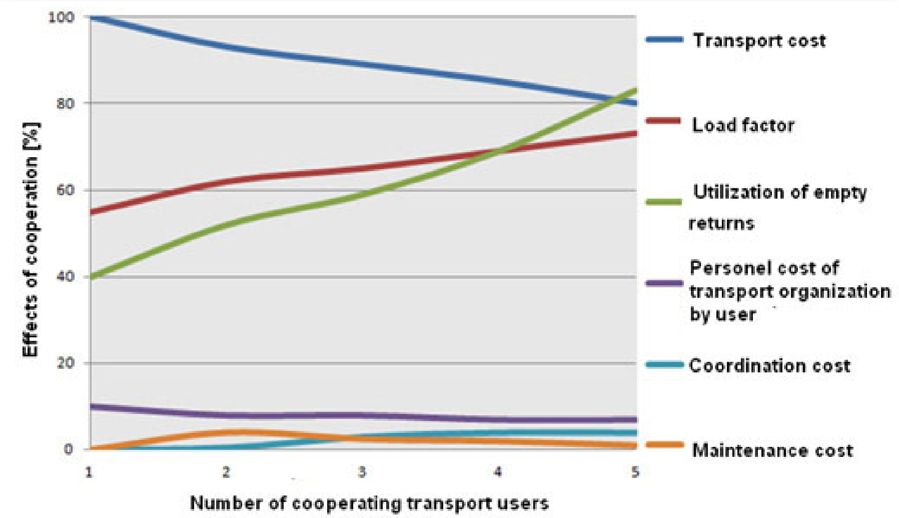


Fig. 6. Effects of cooperation between independent groups of companies on the T-Scale platform

4 Conclusions

Presented in the paper cooperative business model for virtual logistics cluster requires sharing knowledge and information along the supply chain, according to the common and easy to understand language – standards for data exchange. To achieve this, the information and communication systems used for managing transport and logistics operation must be interoperable and the actors need to be able to share that information according to their own business rules. One of the major challenges in development of new virtual logistics clusters is how to create solution which would be able to change this situation. Luckily, almost unlimited access to the Internet makes possible cooperation in the area of transport process not only between big company and also SMEs. The lack of consistency of business process performed by particular entities and the variety IT systems used by companies, cause problems with automatic partners networking. The presented by authors approach provides framework for planning and coordination of transport according to the concept of co-modality. The tool helps to form virtual clusters in order to reach common goal of cost effectiveness as well as the sustainable development goals (congestion reduction).

References

1. Porter, M.: Clusters and the New Economics of Competition. Harvard Business Review, 77–90 (November-December 1998)
2. Pedersen, T.J., Paganelli, P., Knoors, F.: One Common Framework for Information and Communication Systems in Transport and Logistics. DiSCwise Project Deliverable, Brussels (2010)
3. Golinska, P., Hajdul, M.: Multi-agent Coordination Mechanism of Virtual Supply Chain. In: O'Shea, J., Nguyen, N.T., Crockett, K., Howlett, R.J., Jain, L.C. (eds.) KES-AMSTA 2011. LNCS, vol. 6682, pp. 620–629. Springer, Heidelberg (2011)
4. Hajdul, M.: Model of coordination of transport processes according to the concept of sustainable development. LogForum 3(21), 45–55 (2010)
5. GS1 standards in transport and logistics, GS1 Global Office, Brussels (2010)

Supply Chain Configuration in High-Tech Networks

Arkadiusz Kawa and Milena Ratajczak-Mrozek

Poznań University of Economics,
al. Niepodleglości 10, 61-875 Poznan, Poland
{arkadiusz.kawa,milena.ratajczak}@ue.poznan.pl

Abstract. The supply chain configuration has recently been one of the key elements of supply chain management. The complexity of the relations and variety of the aims of their particular members cause it to be very difficult to build a supply chain effectively. Therefore, finding a feasible configuration in which both the business network and the company can achieve the highest possible level of performance constitutes a problem. Authors proposed the SCtechNet model based on graph theory, business network concept and the competitiveness indicator that helps to solve this problem by dynamic configuration of supply chains. The simulation results based on proposed model are presented and discussed.

Keywords: business network concept, supply chain configuration, NetLogo.

1 Introduction

In the face of globalization and the development of the knowledge-based economy, the question arises if changes in the global market environment contribute to the creation of new potential determinants of company competitiveness. Such new determinants may be network relationships and business networks. Firms may find participation in network relationships and business networks an essential determinant for developing competitiveness and the ability to create and use such relationships may become a necessity that will ensure their success.

According to the network approach consistent with the main IMP Group research [12] thread, a business network constitutes a collection of long-term formal and informal relationships (direct and indirect) which exist between two or more entities [8]. Within the considered framework, a system of links often is characterized as being decentralized and informal.

The principle of strategic equality of business entities diverges a great deal from economic reality. Often, it is possible to identify a firm(s) which plays a dominant role within the framework of linked entities in this respect. Firms, with increasing frequency, consciously create business networks concentrated around them. These types of relationships illustrate the strategic approach of the development of network links. The strategic approach [3, 9] stresses the active and conscious development of a network of relations and the presence of one main entity (flagship company) intentionally building a strategic network. The main characteristic of relations between the partners of a network is the asymmetric and strategic control exercised by

J.-S. Pan, S.-M. Chen, N.T. Nguyen (Eds.): ACIIDS 2012, Part I, LNAI 7196, pp. 459–468, 2012.

one flagship company over the remaining (independent or “slightly” dependent) firms. Within partner firms, the only activities which remain autonomous are those which are not included in the cooperation network – the flagship company only has strategic control over those aspects of its partners’ business systems which are dedicated to the network.

The ideology of a strategic network can be linked to the concept of logistics networks and supply chain. The supply chain can be defined as a network of connected and interdependent organizations which acting on the basis of joint cooperation, jointly control, manage and improve resource and goods flows from suppliers to the final consumer. Therefore, supply chain management constitutes a fragment of the overall subject of business networks and this perspective will be assumed in this article.

Despite the fact that the significance of cooperation and relationships between companies is emphasized within research [1, 6], the network approach is used to a relatively limited extent to comprehensively analyze the process of building company competitiveness. Previous studies relating to the impact of network relationships and business networks on competitive advantage are very fragmented. Attention is devoted to the isolated effects exerted by network relationships (e.g. creation of trust, the positive impact of relationships on the learning process and the resulting competitiveness), there is, however, a lack of comprehensive research devoted to the holistic quantification of the mechanism of building company’s competitiveness with the application of network relationships and business networks.

Business networks and network relationships require a new look at the issue of a company’s competitiveness which has to take in to account new quality of relationships connecting business entities The network approach accents the importance of long-term relations between entities which are to lead towards measureable benefits for the whole network as well as individual firms. There is, however, a lack of in-depth analysis of short-term relations from a network perspective. Meanwhile, short-term relations often occur and it appears that they can also result in measureable benefits for firms.

2 Previous Work

This article continues the research into the development of competitive advantage and competitiveness among high-tech firms as a result of network relationships presented at [11]. An empirical analysis of 74 Polish firms was carried out during the research. The results confirmed that vertical links are more often developed among Polish high-tech firms (buyers and suppliers) than horizontal links (between competitors or institutions of research or education). Generally, firms continue to cooperate with each other within the framework of the supply chain increasingly often. At the same time, high-tech firms with an identified competitive advantage (the better firms) more intensively utilized the relations established within the supply chain.

A firm’s competitive advantage is treated as a relative measure of the quality of a firm’s operations and is defined through the prism of the relative differences in financial and non-financial performance (performance differentials) with respect to the achievements of the closest competitors. Competitive advantage is studied based on a consolidated formula including total income, total sale, return on investment

(ROI), and market share. Due to difficulties in comparing companies of various sizes or operating in different markets, a subjective method of assessing their activities in comparison with their closest competitors that is based on a relative assessment of the enterprises themselves was adopted. The 5-point Likert scale was used for the assessment. The respondents, by answering the questions posed in the questionnaire relating to four of the effects of performance (total income, total sale, ROI, market share) were to provide their own self-assessment in relation to the closest competitors (5 – considerably worse, 4 – worse, 3 – almost the same, 2 – better, 1 – considerably better)[1]. The application of such an evaluation method facilitates the comparison of results with those of other firms with different business characteristics. The adaptation of this evaluation method is based upon the earlier experiences of Fonfara [5].

The analysis of firms' performance in four areas with respect to the closest competition served to construct a competitiveness indicator (CI). This indicator was defined as the average of the four results – overall profit, market share, sales volume and ROI. In the subsequently presented deliberations, the lower the positive deviation of the indicator from the value of 3 (remembering that in the adopted scale, the value of 3 indicated "almost the same") the greater the advantage.

The statistical analysis which was carried out in two stages confirmed the adopted approach and the justification for linking the four defined elements of a firm's results in to one indicator.

3 SCtechNet Model

The choice of contractors for the transaction is usually guided by the criteria of resources availability and price competitiveness. Typically, an enterprise inquires several subjects individually and then chooses the most attractive offer. Unfortunately, it is a very time-consuming task which requires efficient data interchange and analysis. Therefore, there is a need for solutions which will make it possible to currently browse, analyze and choose the best of all available offers meeting predefined requirements in a short time. One such solution is definitely the SCtechNet model which allows a data flow within the whole business network.

The SCtechNet model (Supply Chain Configuration in High-Tech Networks) is a development of DyConSC model [7, 10] and it is extended here with the business network concept and the competitiveness indicator. It is mainly aimed at building dynamic and flexible temporary supply chains within the business network. The SCtechNet enables each entity of the supply chain to independently adjust their plans in such a way that they become optimal both within one high-tech enterprise and the whole supply chain.

This model is based on service-oriented architecture (which grants system interoperability) and the graph theory. All IT systems of enterprises are interconnected and enable software agents to access their content. Autonomous agents are capable of identifying which tasks and customers' requirements can be satisfied. Agents representing various enterprises from particular tiers cooperate with one another, coordinate and negotiate conditions to achieve the common goal whereas every agent

[1] For the purposes of this article, the scale has been reversed.

may try to attain the target of the individual user delegating it. Thanks to that, optimizing supply chains holistically and obtaining significant benefits for the final customer become attainable, even if supply chain partners have differing objectives, perspectives and processes [2, 7].

In the SCtechNet model, four tiers of enterprises have been distinguished. The first tier is represented by high-tech flagship company (FC) followed by assemblers, suppliers, and factories. For example, assemblers build up printers, suppliers provide subassemblies, and factories cater raw materials. FC initiates the configuration of supply chain which is caused by the final customer. So, supply chains are created for the needs of a specific transaction evoked by the customer's demand. FC manages the whole supply chain process of a product in real time, from the receipt of the order through gaining resources necessary for the production to the delivery of ready products to the customer. FC is also engaged in the optimization of the already existing supply chains and the control of their efficient accomplishment so that the customers' expectations related to service quality are met and the costs are reduced at the same time. However, the remaining enterprises from the network are directly responsible for the organization and co-ordination of the streams (of goods and information) generated by the suppliers and recipients of the next tier.

The high-tech business network presented above comes in the form of a stratified, directed graph consisting of n sources and one sink. It is noteworthy, though, that figure 1 deliberately does not show horizontal relations (between suppliers of the tier). It has been assumed that the only existing relations are the vertical ones (between suppliers and receivers).

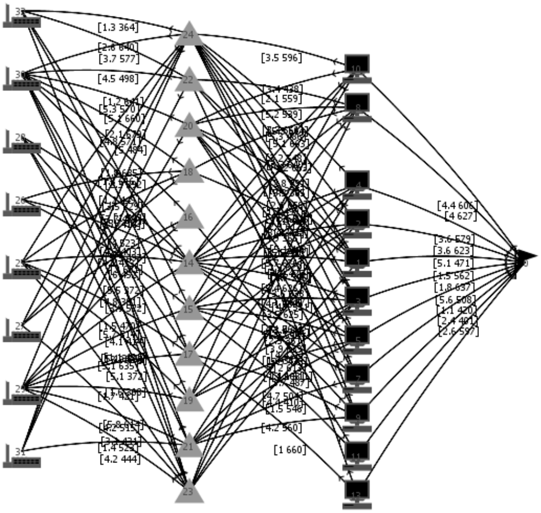


Fig. 1. Screenshot of exemplified high-tech business network in the NetLogo platform

Among the subsequent tiers a flow of goods and information about them takes place. All supplies are conducted sequentially so no tier can be omitted. A flow (edges, according to graph theory) of goods in certain quantities takes place between the entities (nodes) in business network. The criteria for the choice for the preceding entity comprise the price of products or services and the competitiveness indicator (CI). The first criterion is a precondition. If the price meets the requirements, CI is comparable in the next step. CI facilitates finding the best flow (of the lowest values) with an appropriate capacity. It boils down to the minimum cost flow problem, included in the aforementioned graph theory and described below.

Let us assume that in network G (enterprise network) a flow of value θ from s (source, i.e. factory) to t (sink, i.e. FC) is sought. Value θ represents the total demand of FC. The costs d_{ij} of sending a flow unit along an edge (i, j) and maximum capacities c_{ij} of particular edges are defined, too. Additionally, let us mark the flow f_{ij} over edges (i, j). The notion of minimum cost flow may be formulated in the following way [4]:

The sum must be minimized $\sum_{(i,j)} d_{ij} f_{ij} \rightarrow min,$

The assumptions are as follows:

1. For each edge (i, j) in network G
   $0 \leq f_{ij} \leq c_{ij},$
2. For node (entity) s
   $\sum_i f_{si} - \sum_{i,} f_{is} = \theta,$
   where in $\sum_i f_{ij}$ all edges entering node j are summed and in $\sum_j f_{ij}$ all edges going out of node i are summed.
3. For node (entity) t
   $\sum_i f_{ti} - \sum_i f_{it} = -\theta,$
4. For the remaining nodes (entities) j
   $\sum_i f_{ji} - \sum_i f_{ij} = 0.$

Although the model presented above describes the task of linear programming, solving it by general liner programming methods is ineffective due its network structure. In this case the Busacker-Gowen (BG) algorithm is helpful which is presented in [4]. This method consists in increasing the flow along consecutive paths augmenting as much as their capacity allows. The order of appointing paths depends on their length which, in this case, is determined by unit costs. If the flow has achieved value θ, computing finishes. Otherwise, the network is modified in such a way that the flow found so far is taken into account. In the residual network (G*) the cheapest path from s to t needs to be found and the greatest number of units is send along it. These two stages are repeated until the flow of the predefined value θ is accomplished or until the current network contains the path from s to t.

In order to find the cheapest chain from the source to the sink the algorithm of finding the shortest paths must be applied. The SCtechNet model has used the BMEP (Bellman, Moore, d'Escopo, Pape) algorithm (see more in [4]).

4 Implementation and Simulation Experiments

The SCtechNet model was implemented in the NetLogo platform. This platform allows both to model and conduct simulations. In the simulating model, four kinds of "breeds" were distinguished: flagships, assemblers, suppliers, and factories. Thanks to that, different behaviors and "agent sets" of those breeds can be defined.

The arrangement of the network is randomly generated. In the first step, the requirements for the price are checked, i.e. the level of price cannot be higher than 30 euro per unit. If the nodes do not have any corresponding connections (because the condition is not met) with their predecessors and/or sequential constituents, they are eliminated from the network. In this case the link between them is not created. However, it does not mean that those nodes will not be taken into account in other business networks.

There is only one flagship company in each network. It is worth to emphasize that different FCs may compete with others from other business networks. The quantity of entities in the second, third and fourth tier is assumed to range from 10 to 100. This number can be increased or decreased with the slider (nodes number). The entities are represented by agents which interact and communicate, e.g. request current production capability, product price, product quantity etc.

Other parameters (increased or decreased with the sliders, too) are as follows:

_ SCD (supply chain demand) – demand of the flagship company which equals the whole supply chain demand (it changes from 1 000 to 10 000 units);
_ SI (supply indicator) – factor of supply changeability of particular entities excluding flagship company (it changes from 20% to 100%).

The properties of link agents between constituents were chosen randomly as a pair of competitiveness indicator (CI) and capacity (CA). The first one is between 1.0 and 5.0. This indicator has been described above (cf. section 2). The CA is established as a variable according to the following procedure: SCD * SI + random (SCD * SI + 1). For example, if SCD = 50 and SI = 0.5, then CA amounts to not less than 25 and not more than 50.

The FC demand can be completely or partially satisfied. It depends on the quantity supply of the preceding constituents. Similarly, the demand of enterprises from the 2nd tier can be fulfilled completely or partially and also depends on the supply of the entities from the 3rd tier, and so on.

In order to find the best supply chain (so the shortest path in the graph), the BG and BMEP algorithms are activated (cf. Section 3). In the network, there can be at least one such chain. Their number depends on the supply and demand changeability.

The main aim of the simulations carried out was to study how the changes of the number of nodes and the supply indicator influence the supply chains number and the average level of the competitiveness indicator along the business network.

In the first simulations, the number of entities in a particular tier was changing from 10 to 100, incrementing each time by 10 (simultaneously, this number in other tiers was stable and equaled 20) on the assumption that SCD = 10 000 and SI = 20%. They were run 1 000 times for each case which gives a total of 10 000 times.

The findings of the simulations show that as the nodes number augments, the average number changes (see fig. 2). For example, if the assemblers number increases 10 times (from 10 to 100 nodes), the average supply chains number decreases only by 12%, but in the case of factories it rises by 3%. It may, then, be concluded that the average number of supply chains is not considerably influenced by the change of the number of nodes in particular tiers.

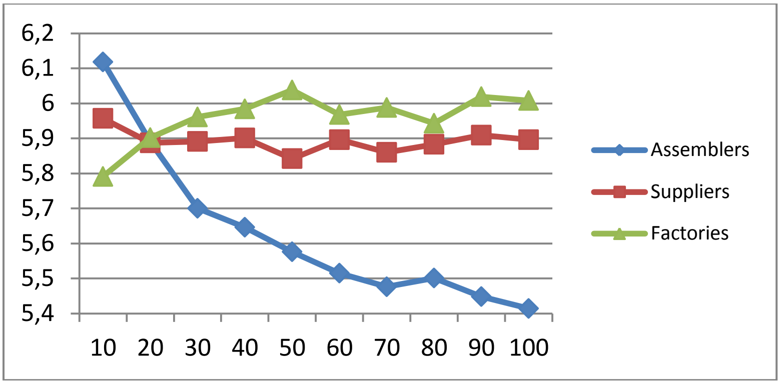


Fig. 2. Influence of enterprises number change in particular tiers on supply chains number in the network

It is different, though, in the case of the average level of the whole competitiveness indicator (CI), in which changes are more visible (see fig. 3). When there is an increase in the nodes number, CI decreases by 15% on average, but for assemblers the fall is greatest and reaches 24%. Thus, the growth of the number of nodes in particular tiers causes the average CI along a supply chain to diminish. It can be explained by the following dependence: the more suppliers there are in a given tier, the higher the competitiveness among them is and the better the conditions become for the final customers. The CI of each tier tends towards 1.0, so the CI of the whole supply chain tends towards 3.0. Figure 3 shows one more correlation – the greater the nodes number in the vicinity of the flagship company (so assemblers, suppliers, and factories successively), the lower the CI.

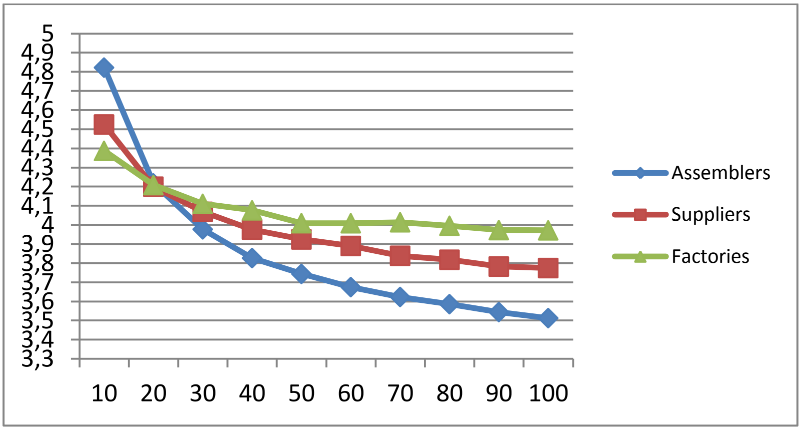


Fig. 3. Influence of enterprises number change in particular tiers on the average competitiveness indicator in the business network

Next, the impact of the variation of the supply indicator (so the capacity of enterprises) on the supply chains number and the competitiveness indicator was studied, on the assumption that the nodes number is stable and amounts to 20 and SCD = 10 000. In the simulations, the variable supply indicator was shifted by 20 percentage points, from 20% to 100%. The simulations carried out show that the supply chain number falls sharply from 13 to 1, i.e. about 92% (see fig. 4). It is interesting that the supply indicator raised twice (i.e. from 10% to 20%) causes the average supply chains numbers, which can satisfy the FC demand more quickly, to drop from 13 to 6, i.e. 53%. As a result of the increase of the supply indicator from 10% to 100%, CI declines by 22% (see the red line in fig. 5). The greatest drop (about 11%) takes place in the first step (from 10% to 20%). It may be stated on the basis of the analysis of the data from this simulation experiment that it is more profitable to collaborate with enterprises which have greater capacities and can offer greater supply (assuming that the nodes number is stable). It reduces the number of supply chains.

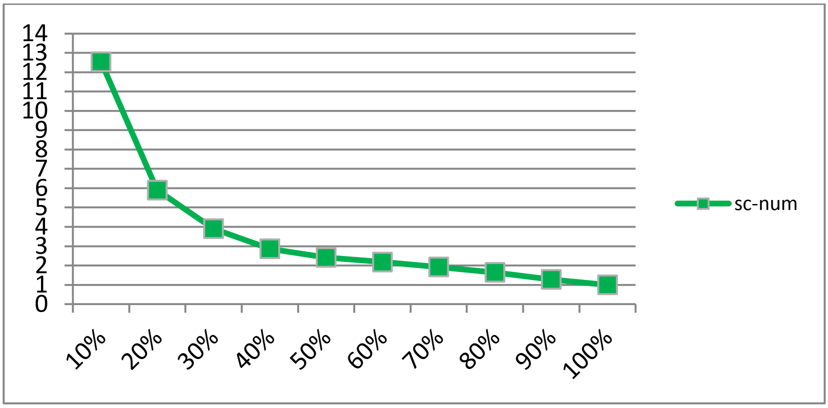


Fig. 4. Influence of factor of supply changeability on the supply chains number in the business network

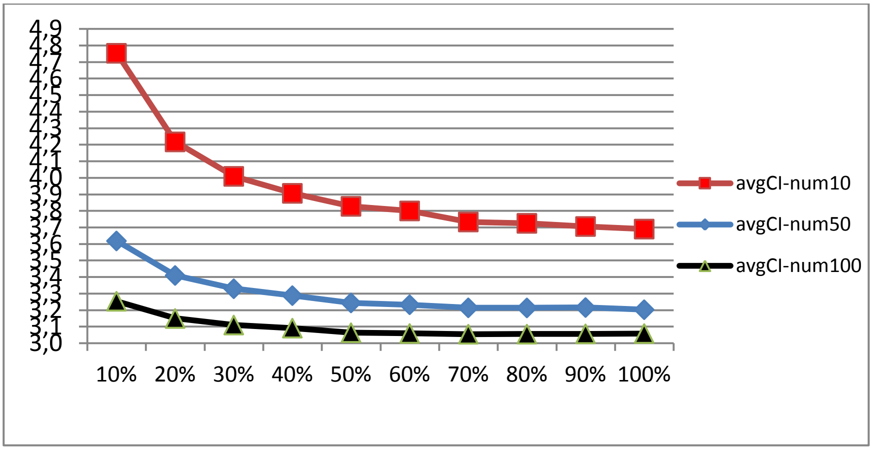


Fig. 5. Influence of factor of supply changeability on the average competitiveness indicator in the business network

In the last part of the simulations, the impact of the variation of the supply indicator on CI was investigated, having presumed that the nodes number is stable and amounts to 50 (see the blue line in fig. 5), and then to 100 (the black line). Here, the correlations are the same as above – the growth of the number of nodes in tiers causes the average CI along a supply chain to diminish.

5 Conclusion

Highly innovative, R&D intensive high-tech companies need solutions aiming at improving their supply chain, rising their competitiveness and reducing costs. Application of the agent technology and graph theory in logistics allows departing from fixed supply chains, in which enterprises are dependent on one another, and replace them with dynamic configurable supply chains. Including business network concept and competitiveness indicator additionally extends the analysis.

The proposed SCtechNet model offers many benefits for the network of enterprises, its participants and the final customer. Some of the most important ones have been distinguished below: finding solutions rising competitiveness of all network entities individually and supply chain holistically, giving answer to the question how to build dynamic and flexible temporary supply chains within the business network and with whom it is more profitable to collaborate with. The model allows also to currently browse, analyze and choose the best of all available offers meeting predefined requirements in a short time. It gives the possibility to build scenarios and carry out simulations independently.

Acknowledgements. The paper was written with financial support from the Foundation for Polish Science [Fundacja na rzecz Nauki Polskiej].

References

1. Campbell, A.J., Wilson, D.T.: Managed Networks. In: Iacobucci, E. (ed.) Networks in Marketing. Sage Publications, USA (1996)
2. Chiu, M., Lin, G.: Collaborative supply chain planning using the artificial neural network approach. Journal of Manufacturing Technology Management 15(8) (2004)
3. D'Cruz, J.R., Rugman, A.M.: Developing international competitiveness: the five partners model. Business Quarterly 8(2) (1993)
4. Deo, N., Kowalik, J.S., Sysło, M.M.: Discrete Optimization Algorithms with Pascal Programs. Prentice-Hall Inc., Englewood Cliffs (1983)
5. Fonfara, K.: A typology of company behaviour in the internationalisation process (a network approach). In: 24th IMP-Conference in Uppsala, Sweden (2008)
6. Ford, D., Gadde, L.E., Håkansson, H., Snehota, I.: Managing business relationships. Wiley (2003)
7. Fuks, K., Kawa, A., Wieczerzycki, W.: Dynamic Configuration and Management of e-Supply Chains Based on Internet Public Registries Visited by Clusters of Software Agents. In: Mařík, V., Vyatkin, V., Colombo, A.W. (eds.) HoloMAS 2007. LNCS (LNAI), vol. 4659, pp. 281–292. Springer, Heidelberg (2007)

8. Håkansson, H., Snehota, I.: No business in an island: the network concept of business strategy. Scandinavian Journal of Management 5(3) (1989)
9. Jarillo, J.C.: Strategic networks. Creating the bordless organization, Butterworth Heinemann (1995)
10. Kawa, A.: Simulation of Dynamic Supply Chain Configuration Based on Software Agents and Graph Theory. In: Omatu, S., Rocha, M.P., Bravo, J., Fernández, F., Corchado, E., Bustillo, A., Corchado, J.M., et al. (eds.) IWANN 2009, Part II. LNCS, vol. 5518, pp. 346–349. Springer, Heidelberg (2009)
11. Ratajczak-Mrozek, M.: Sieci biznesowe a przewaga konkurencyjna przedsiębiorstw zaawansowanych technologii na rynkach zagranicznych, Wydawnictwo Uniwersytetu Ekonomicznego w Poznaniu, Poznań (2010)
12. `http://www.impgroup.org`

Identification and Estimation of Factors Influencing Logistic Process Safety in a Network Context with the Use of Grey System Theory

Rafal Mierzwiak, Karolina Werner, and Pawel Pawlewski

Poznan University ofTechnology, Strzelecka 11, 60965 Poznan, Poland
{rafal.mierzwiak,karolina.werner,
pawel.pawlewski}@put.poznan.pl

Abstract. An article presents identification methodologies and estimation of factors which influence a logistic process safety in a network context through the use of a network thinking methodology. A proposed version of a network thinking methodology for the use of a safety analysis of logistic processes realisation, unlike an originalProbst and Urlich concept, only uses its modified stages. A significant element which differentiates a solution presented by the authors is the use of one of grey systems theory's methods so called Grey Relational Analysis in order to quantitatively formulate a common experts' opinion concerning an impact mutual force of identified factors. Knowledge of a correlation force between factors is a basis to classifying factors into groups which are used for taking appropriate and optimal managing actions.

A solution presented in the article will be tested with reference to real data for a buying decision process in a company of a chemical branch using a computer support that simplifies computation processes.

Keywords: logistic process safety, network thinking methodology, Grey Relational Analysis.

1 Theoretical Bases of Problematic Aspects of Logistic Process Realization

Safety is a state that gives a mood of confidence and assures its preservation and opportunities for improvement [11]. Such understood safety in management can be referred to many aspects among which there is also processes realisation safety. Here a process is treated as a sequential series of actions that starts from gaining and processing informational and material resources. After transformation, which is specific for a given process, they become resources for subsequent series of actions realised by internal and external customers [9].

A specific group of processes which is sensitive to risks that lower a level of their safety are logistic processes. A process is named a logistic process when an arrangement, a state, and a flow of its constituents require coordination with other processes due to criteria of location, time, costs, and effectiveness of fulfilling company's superior goals [6]. Typical logistic processes include processes of supply,

J.-S. Pan, S.-M. Chen, N.T. Nguyen (Eds.): ACIIDS 2012, Part I, LNAI 7196, pp. 469–477, 2012.

storage, transport, distribution, inventory management, and packaging management. From the viewpoint of every production company, a supply process (also called a buying decision process) is particularly important as ensuring production continuity and hence a sale in a company depend on its arrangement.

Sale continuity influences a company's profit. On the other hand, a supply process is a process which generates costs in a company. Assuming that a goal of business activity is to maximise incomes and minimise expenditures, thus, the performers of a supply process face a huge challenge. On the one hand, they have to supply a company with goods necessary for production complying with qualitative requirements in order to satisfy mass market's needs. On the other hand, they have to minimise supply costs (prices of resources, transport costs, storage costs) in order to make a profit. Appropriate balance between those aspects is strongly connected with an analysis of processes realisation considering a safety criterion and proper risk optimising.

2 Network Modeling of Logistics Process Safety

A network modelling starts with the premise about a multi-perspective approach to solve analysed problems through the participation of many stakeholders who represent different viewpoints and requirements resulting from the mentioned viewpoints, and perceived factors which influence the way of fulfilling those requirements [14 p. 107].

An analysis of stakeholders as a starting point in a network modelling must include a large amount of perspectives within which a company is working and reflect strategic, tactical, and operational spheres. The above postulate should be included in a choice of a group responsible for conducting the whole process of a network modelling. A properly selected group guarantees that further stages of the analysis are correctly carried out which means that requirements according to an analysed process will be accurately formulated and important factors related to those requirements will be accurately chosen. Factors determined in this way are further used for constructing a network of relations between factors. An example of such a network is graphically presented in Figure 1.

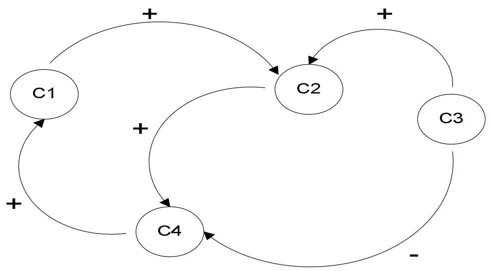


Fig. 1. Network relations presented in a graphical way

In Figure 1, a direction of arrows means a direction of interaction of one factor on another one. Symbols + and – mean a character of this interaction, namely if an influence of one factor on another one is just like the fact that the growth of one factor's value causes the growth of another factor's value, then a character of such an interaction is marked by + symbol and called a positive interaction. In the opposite case, such an interaction is marked by – symbol and called a negative interaction.

It is essential to know a direction of factors' interaction and a character of their interaction. However, these are not all the information necessary to profound knowledge of properties of a constructed network. Firstly, knowledge about a strength which factors use to interact between themselves is also needed. Secondly, questions "*In what way can we influence the factors?*" and "*Are the factors susceptible to this influence (what can be determined by the use of a term of controllability)?*" should be answered. To this aim, experts' opinions need to be referred to due to a variety of existed factors and a difficulty of gaining such an information considering costs and measuring obstacles. Therefore, the authors propose to use a special algorithm to determine experts' common opinion through using Grey Relational Analysis which is one of Grey Systems Theory's techniques.

Grey Systems Theory was proposed for the first time in 1982 by a Chinese scientist Julong Deng to model phenomena in which data are uncertain and incomplete and mechanisms which rule modelled phenomena are only partially known. The theory has remained unknown for a long period of time due to the fact that its first systematic contribution in English language appeared just in 1989 [4] and the first English manual was published in Europe just in 2005. Despite the difficulties in popularisation of Grey Systems Theory, it has finally got many interesting applications especially in technical and economic sciences [1], [2], [3], [5], [13].

3 A Study of Strength of Interaction between Factors and Their Controllability's Estimation

A study of strength of relation between factors with the use of experts' estimation and a procedure of Grey Relational Analysis starts from defining a scale on which the estimation will be done. A proposed solution is accepting a five-point scale where 1 means a very little strength of relation between factors whereas 5 is a very big and significant strength of interaction. After accepting the scale, each of the experts performs the estimation for every identified pair of factors among which a relation resulting from a graphic representation of a network takes place. Mathematically, this can be written in the following way

$$E\left(c_i \rightarrow c_j\right) = \{E(1), E(2) \dots E(n)\} i, j \in (1, m) \tag{1}$$

where,

$E\left(c_i \rightarrow c_j\right)$ - a vector of assessments done by experts for a factorc_i – interacting on a factor c_j

$E(1), E(2) \dots \dots E(n)$ – an estimation of nth factor concerning strength of relation between factors $c_i$and c_j

n – a number of experts,
m – a number of identified factors
i, j - numerical designation of factors which undergo estimation

Having an expert's opinion, we construct a fictional vector of ideal estimations which will include homogenous elements of a value that equals 1. The mentioned vector is as follows

$$E_{IDEAL} = \{ E_{IDEAL}(1) = 10, E_{IDEAL}(2) = 10, \dots E_{IDEAL}(n) = 10\} \quad (2)$$

Having an experts' opinion vector $E\left(c_i \rightarrow c_j\right)$ and an ideal vector E_{IDEAL}, we can begin Grey Relational Analysis procedure which aim is to agree on the experts' opinion by calculating a grey relation grade. Its result should be in the span from 0 to 1. However, when the value is closer to 1, then a higher strength of relation between analysed factors $c_i \rightarrow c_j$can be ascertained.

The first step leading to calculating a grey relation grade between $c_i \rightarrow c_j$is to expose a vector to an average operator$E\left(c_i \rightarrow c_j\right)$. A vector's operation is presented below

$$E'\left(c_i \rightarrow c_j\right) = \{E^{'}(1),^{'}E(2) \dots E^{'}(n)\} \quad (3)$$

where:

$$E^{'}(n) = {}^{E(k)}\Big/_{\frac{1}{n}\sum_{k=1}^{n} E(k)} \quad k \in (1, n) \quad (4)$$

$E^{'}(n)$ – a value of n^th expert's estimation after its exposition to an average operator
$E(n)$ – an estimation given by n^th expert
$\frac{1}{n}\sum_1^n E(n)$- experts' estimations average

After receiving a vector $E'\left(c_i \rightarrow c_j\right)$, the next step is to calculate a vector Δpresented in formula 5.

$$\Delta = \{ \Delta(1), \Delta(2) \dots \Delta(n)\} \quad (5)$$

where

$$\Delta(k) = |E'(k) - E_{IDEAL}(k)|, k \in (1, n) \quad (6)$$

In the succeeding step, M and m are reckoned according to formulas (7) and (8).

$$M = \max \Delta(\, k) \quad (9)$$

$$m = min\Delta(k) \quad (10)$$

Values received earlier allow to creating a vector of grey relation grades for kthitem in the following form

$$\gamma = \{\gamma_1, \gamma_2 \cdots \gamma_n\} \tag{11}$$

where

$$\gamma_k = \frac{m+\rho M}{\Delta_k+\rho M} k \in (1,n) i \rho \in (0,1) \tag{12}$$

γ_k – a grey relation grade for kth item

ρ – a grade selected individually to an analysed case which usually equals 0,5 in applications available in subject literature

When having grey relation grades for kth item calculated, we can calculate an average grey relation grade that projects strength of relation between c_i, c_jaccording to a formula presented below

$$\Gamma_{c_i \to c_j} = \frac{1}{n}\sum_{k=1}^{n} \gamma_k * \frac{1}{n}\sum_{k=1}^{n} E(k) \tag{13}$$

When calculating $\Gamma_{c_i \to c_j}$for each pair of factors and assuming that $\Gamma_{c_i \to c_j} = 0$for factors for which according to a constructed network it appears that there is a lack of interactions, we can construct a matrix in the form of the following formula

$$A = \left[\Gamma_{c_i \to c_j}\right]_{ij} \tag{14}$$

Having at disposal a matrix A, we calculate for it two values $\bar{R}_i$ and $\bar{R}_j$which are an average of values appearing in the matrix's verses and columns respectively what can be expressed by formulas (15) and (16).

$$\bar{R}_i = \frac{1}{m}\sum_{i=1}^{m} \Gamma_{c_i \to c_j} \tag{15}$$

$$\bar{R}_j = \frac{1}{m}\sum_{j=1}^{m} \Gamma_{c_i \to c_j} \tag{16}$$

Values $\bar{R}_i$ and $\bar{R}_j$ allow to determining factors' character and on their basis four groups can be differentiated, namely

- active factors when $\bar{R}_i \leq 0{,}5$ and $\bar{R}_j > 0{,}5$. These are factors which have influence on other factors, however, they cannot be influenced themselves.
- lazy factors when $\bar{R}_i \leq 0{,}5$ and $\bar{R}_j \leq 0{,}5$. These are factors which do not influence the other factors but they can be influenced themselves.
- passive factors when $\bar{R}_j \leq 0{,}5$ and $\bar{R}_i > 0{,}5$. These are factors which do not undergo the influence of other factors, however, they cannot influence other factors themselves.

- crucial factors when $\bar{R}_i > 0{,}5$ and $\bar{R}_j > 0{,}5$. These are factors which have a strong influence on the other factors but simultaneously they can be strongly influenced by those factors.

A very important issue in a factors' analysis which are included in a given network is their estimation according to a criterion of susceptibility to controllability. Accepting the criterion of susceptibility to control, two groups of factors can be distinguished, namely manageable factors and non-manageable factors. The former are the factors which managers have influence on. The latter cannot be changed by managers. A selection procedure of the mentioned groups is similar to a procedure of determining active, passive, crucial, and lazy factors. The difference lies in the fact that the experts do not estimate strength of relation but a possibility to control these factors. This estimation is also done on a ten-point scale. This way, a vector of a given factor's controllability's estimations is received. On this basis, using formulas from (3) to (13) with a little formal corrections connected with symbols in superscripts we determine Γ_{c_i} grade. When we have the mentioned grade, we can claim that a factor is non-manageable if $\Gamma_{c_i} \leq 0{,}5$. However, a factor is considered to be manageable when $\Gamma_{c_i} > 0{,}5$.

4 Testing a Developed Methodology on a Buying Decision Process in a Chemical Branch

In order to test a methodological approach presented above a chemical company was chosen. In the first step, a knowledge how a buyer decision process goes is gained according to thirteen quality management system's procedures and interviews with company's employees. In a result of this analysis, a team that includes five experts was formed. The team determined groups of stakeholders in a described process and their requirements connected with a safety of a buyer decision process. The outcomes received in this stage are presented in Table 1.

Table 1. Stakeholders in a buying decision process and their requirements concerning safety of a process realisation

	Stakeholders' Name	Stakeholders' requirements
1.	Logistics Specialist	To supply a company in appropriate goods in exact quantity and quality, and in right time and place at appropriate price according to buying decision process's procedures.
2.	Head of Logistic Department	To supply a company in appropriate goods in exact quantity and quality, and in right time and place at appropriate price according to buying decision process's procedures.
3.	Deputy Head of Logistic Department	To supply a company in appropriate goods in exact quantity and quality, and in right time and place at appropriate price according to buying decision process's procedures.

Table 1. (*continued*)

4.	Marketing and Sales Director	To realise a buying decision process according to a plan and procedures.
5.	Employees in TP1 Department Employees in TP2 Department	To receive exact goods in exact quantity and quality in the right time and place. To take and store a delivery according to procedures.
6.	Quality service employees	To examine delivery according to quality standards and requirements following the procedures.
7.	Board of Management Representative concerning Quality Management System	A buying decision process must proceed according to obligatory procedures.
8.	Employees in Research and Development Department	Delivering new resources needed to research.
9.	Employees in Financial Department	Receiving correctly formulated invoices in order to pay for deliveries on time.
10.	Employees in Accounting Department	Receiving paid invoices for deliveries on time and accounting them.
11.	Financial and Accounting Director	Supplying a company in exact goods according to quality standards at a minimal price and at a minimal transport cost.
12.	Chairperson of the Company	To guarantee smoother production at a minimal cost through resources quality coherence.
13.	Company's customer	To receive exact goods in exact quantity and quality in the right time and place at a minimal price.

On the basis of stakeholders' requirements, the experts determined 19 factors influencing a logistic process safety in significant way and these are c_1 - resource's availability on the market, c_2 – prices of resources, c_3– quality of resources, c_4 - competitors on suppliers' market, c_5 – exchange rates, c_6 – relations with a carrier, c_7 – legal articles concerning purchase, transport, import of resources, c_8 – employees qualifications who purchase goods, c_9 – professional experience of employees who purchase goods, c_{10} - knowledge of foreign languages by employees who purchase goods, c_{11} – interpersonal features of employees who purchase goods, c_{12} – standardisation of a buying decision process, c_{13} – honesty and reliability of suppliers, c_{14} – information transfer, c_{15}– time of placing an order at a supplier, c_{16} – correctness of a placed order, c_{17} – company's cash flow, c_{18} – honesty and reliability of a carrier, and c_{19} – a king of used mean of external transport.

Having factors selected, the experts developed a network in which directions of factors' interactions and a kind of interaction were determined. On the basis of factors' network, the experts estimated the strength of interaction between factors. The outcomes were estimated on five-point scale where 1 means a very weak strength of interaction and 5 means a very strong interaction. This estimation was used for calculating a grade of strength of interaction $\Gamma_{c_i \to c_j}$between factors. To this aim, an authorial software ROgraph was used. A grade $\Gamma_{c_i \to c_j}$was used for constructing a matrix $A = \left[\Gamma_{c_i \to c_j}\right]_{ij}$, calculating grades $\bar{R_i} = \frac{1}{m}\sum_{i=1}^{m}\Gamma_{c_i \to c_j}$ and

$\bar{R}_j = \frac{1}{m}\sum_{j=1}^{m}\Gamma_{c_i\to c_j}$, and as a result selecting the factors into the following groups of factors, namely active, passive, lazy, and crucial. Final outcomes of calculations performed on the matrix $A$are presented in Table 2.

On the basis of an analysis of results included in Table 2, we can state that all the factors belong to a group of lazy factors which do not cause changes in other factors and simultaneously are not a subject to change. Such situations can be interpreted in the following way, namely a safety of an order process in the discussed company indicates high stability and resistance to distractions what is undoubtedly a big advantage in a normal functioning. However, any attempts of improvements or changes enforced by the market meet resistance. Due to a reason that all the factors which influence a buying decision process are lazy, the authors deemed that it is pointless to examine their controllability.

Table 2. Strength of interaction between factors

i->j		c_j																			
		1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**13**	**14**	**15**	**16**	**17**	**18**	**19**	$\bar{R}_j$
c_j	**1**	**0,000**	1,000	0,579	1,000	0,000	0,000	0,073	0,000	0,108	0,000	0,000	0,125	0,656	0,219	0,230	0,243	0,198	0,036	0,248	**0,248**
	2	0,603	**0,000**	0,864	0,412	0,000	0,000	0,283	0,000	0,097	0,000	0,000	0,000	0,305	0,250	0,271	0,206	0,251	0,195	0,213	**0,208**
	3	0,310	0,603	**0,000**	0,564	0,000	0,144	0,223	0,259	0,000	0,097	0,097	0,000	0,207	0,310	0,259	0,355	0,000	0,304	0,428	**0,219**
	4	0,549	0,838	0,549	**0,000**	0,000	0,149	0,187	0,093	0,219	0,073	0,093	0,115	0,472	0,149	0,187	0,149	0,304	0,110	0,149	**0,231**
	5	0,422	0,420	0,259	0,223	**0,000**	0,000	0,036	0,000	0,000	0,072	0,000	0,097	0,206	0,097	0,073	0,097	0,373	0,073	0,149	**0,137**
	6	0,149	0,125	0,163	0,125	0,000	**0,000**	0,000	0,093	0,149	0,093	0,093	0,073	0,073	0,219	0,110	0,073	0,115	0,373	0,562	**0,136**
	7	0,334	0,000	0,213	0,223	0,000	0,219	**0,000**	0,000	0,149	0,310	0,109	0,213	0,149	0,115	0,093	0,213	0,149	0,149	0,324	**0,156**
	8	0,072	0,000	0,000	0,097	0,000	0,213	0,110	**0,000**	0,133	0,305	0,213	0,334	0,149	0,320	0,562	0,472	0,364	0,250	0,585	**0,220**
	9	0,187	0,206	0,304	0,310	0,000	0,320	0,000	0,406	**0,000**	0,223	0,213	0,656	0,223	0,320	0,756	0,792	0,364	0,213	0,320	**0,306**
	10	0,320	0,125	0,000	0,109	0,000	0,258	0,000	0,437	0,576	**0,000**	0,149	0,072	0,223	0,223	0,187	0,201	0,201	0,149	0,093	**0,175**
	11	0,000	0,125	0,097	0,115	0,000	0,297	0,000	0,223	0,292	0,250	**0,000**	0,097	0,292	0,473	0,373	0,370	0,334	0,342	0,173	**0,203**
	12	0,249	0,249	0,394	0,405	0,000	0,133	0,144	0,440	0,400	0,000	0,267	**0,000**	0,109	0,324	0,428	0,533	0,036	0,206	0,149	**0,235**
	13	0,459	0,187	0,355	0,403	0,000	0,110	0,149	0,115	0,149	0,036	0,198	0,036	**0,000**	0,206	0,206	0,423	0,576	0,125	0,405	**0,218**
	14	0,310	0,247	0,125	0,247	0,000	0,279	0,093	0,149	0,267	0,133	0,213	0,097	0,324	**0,000**	0,556	0,267	0,231	0,267	0,183	**0,210**
	15	0,389	0,219	0,267	0,462	0,000	0,324	0,231	0,283	0,567	0,072	0,198	0,206	0,230	0,423	**0,000**	0,373	0,251	0,198	0,223	**0,259**
	16	0,223	0,187	0,110	0,403	0,000	0,206	0,173	0,223	0,355	0,198	0,251	0,244	0,428	0,109	0,423	**0,000**	0,173	0,230	0,133	**0,214**
	17	0,545	0,187	0,247	0,163	0,000	0,320	0,000	0,576	0,279	0,149	0,198	0,405	0,316	0,355	0,254	0,109	**0,000**	0,283	0,310	**0,247**
	18	0,223	0,097	0,230	0,163	0,000	0,864	0,110	0,093	0,223	0,036	0,230	0,036	0,036	0,036	0,036	0,036	0,279	**0,000**	0,549	**0,173**
	19	0,269	0,198	0,163	0,223	0,000	0,403	0,230	0,223	0,267	0,115	0,110	0,072	0,149	0,109	0,206	0,206	0,247	0,355	**0,000**	**0,187**
		0,296	**0,264**	**0,259**	**0,297**	**0,000**	**0,223**	**0,108**	**0,190**	**0,223**	**0,114**	**0,139**	**0,152**	**0,239**	**0,224**	**0,274**	**0,269**	**0,234**	**0,203**	**0,273**	

5 Conclusion and Future Research Directions

A formulated method with the use of Grey Systems Theory requires a further development and improvements. A particular emphasis should be put on aspects connected with improving a selection of experts as the authors noticed that when there is a smaller agreement between experts, then a formulated algorithm lowers the results in a significant way. Another important area of a method's improvement should also be an area of classifying the factors into earlier determined groups. A proposal that is worth to consider is to use a grey taxonomic classification i.e. Grey Clustering Analysis or other advanced methods[12],[7].

What is more, a relevant factor which complicates the use of a network thinking methodology in a proposed version is vagueness of used expressions and a lack of terminological coherence. In authors' opinion, this problem can be eliminated by the use of a quality approach to a network's description. The quality approach is developed by Polish scientists within the scope of a general theory which is qualitology. Particularly useful in this area might be the works of Mantura [8] and Szafrański [10] which are unfamiliar due to a lack of translations into English.

References

[1] Akay, D., Atak, M.: Grey prediction with rolling mechanism for electricity demand forecasting of Turkey, Energy, 32 (2007)

[2] Cempel, C.Z.: Zastosowanie teorii szarych systemów do modelowania i prognozowania w diagnostyce maszyn (The use of Grey Systems Theory to model and forecast in machinery diagnostics). Diagnostyka 2(42) (2007)

[3] Chen, C., Ting, S.: A study using the grey system theory to evaluate the importance of various service quality factors. International Journal of Quality and Reliability Management 19(7) (2002)

[4] Deng, J.: Introduction to grey system theory. The Journal of Grey System 1(I), 1–24 (1989)

[5] Hsu, C., Wen, Y.: Improved Grey Prediction Models for Trans-Pacific Air Passenger Market, Transportation. Planning and Technology 22 (1998)

[6] Krawczyk, S.: Logistyka w zarządzaniu marketingiem, p. 38. AE Publishing, Wrocław (1998)

[7] Lin, T.C., Huang, H.C., Liao, B.Y., Pan, J.S.: An Optimized Approach on Applying Genetic Algorithm to Adaptive Cluster Validity Index. International Journal of Computer Sciences and Engineering Systems 1(4), 253–257 (2007)

[8] Mantura, W.: Elementy kwalitologii. Wydawnictwo Politechniki Poznańskiej. Poznań (2011)

[9] Nowosielski, S. (ed.): Procesy i projekty logistyczne, p. 4. UE Publishing, Wrocław (2008)

[10] Szafrański, M.: Skuteczność działań w systemach zarządzania jakością. Wydawnictwo Politechniki Poznańskiej. Poznań (2006)

[11] Szymonik, A.: Bezpieczeństwo w logistyce, Difin, Warszawa, p. 11 (2010)

[12] Tzung-Pei, H., Cheng-Hsi, W.: An Improved Weighted Clustering Algorithm for Determination of Application Nodes in Heterogeneous Sensor Networks. Journal of Information Hiding and Multimedia Signal Processing 2(2), 173–184 (2011)

[13] Wang, T., Liou, M., Hung, H.: Application of Grey Theory on Forecasting the Exchange Rate between TWD and USD. In: International Conference on Business and Information, Academy of Taiwan Information System Research and Hong Kong Baptist University, Hong Kong, July 14-15 (2005)

[14] Zimniewicz, K.: Współ czesne koncepcje i metody zarządzania, PWE, Warszawa (1999)

Implementation of Vendor Managed Inventory Concept by Multi-Dimensionally Versioned Software Agents

Piotr Januszewski and Waldemar Wieczerzycki

Poznan University of Economics, Department of Logistics and Transportation,
al. Niepodległości 10, 61-875 Poznań, Poland
{Piotr.Januszewski,W.Wieczerzycki}@ue.poznan.pl

Abstract. Due to increasingly competitive markets, firms constantly search for new ways to lower their operational costs. One of many areas on which companies focus the most, is inventory management. More and more enterprises lay their faith in complex technical solutions which, in their opinion, will give them competitive advantage over their competitors. This paper is dedicated to demonstrate an innovative way to improve a traditional concept in inventory management (Vendor Managed Inventory), by using agent technologies. We propose an approach to utilize intelligent yet highly mobile software agents which, from an economical point of view, cold be both – cheaper and more effective than traditional information systems.

Keywords: Inventory Management, Vendor Managed Inventory, Intelligent Software Agents, Agent Managed Inventory, Multi-Dimensionally Versioned Agent.

1 Introduction

Inventory management has always been an issue, which companies had to cope with. Nowadays firms tend to constantly search for new ways for decreasing their operational costs, especially those resulting from unnecessary supplies. They try to generate more precise forecasts, establish a better information flow between business partners or optimize their inner processes to ensure a more effective way to manage their stock levels. Moreover firms quickly realized the importance of collaboration within a supply chain and its impact on operational costs. This led to the development of many distribution and inventory control methodologies in management. Those methodologies have been constantly evolving from relatively static to more dynamic models. Good inventory levels are now a measure of business competitiveness [1].

Those concepts were quickly improved by utilizing the potential of information and communication technology (ICT). Simple inventory management applications evolved throughout the years into complex Enterprise Resource Planning (ERP) and Supply Chain Management (SCM) systems. Those complex giants however have some significant cons, which limit their usability for enterprises – they are expensive and their implementation process in extremely time-consuming. Those factors were the key element which inspired the authors to propose a new concept in inventory management systems.

J.-S. Pan, S.-M. Chen, N.T. Nguyen (Eds.): ACIIDS 2012, Part I, LNAI 7196, pp. 478–487, 2012.

In this paper, 2 main goals were achieved. The first goal was to propose an alternate technology to be used in inventory management. The key concept was the usage of agent technologies in the process of inventory management. The main goal of the presented solution is to implement a classic inventory management technique – VMI but in a improved, more effective and broadly available version. The second goal was the demonstration of a new software agent model which is both – intelligent and highly mobile. The proposed agent model is essential to implement the proposed inventory management solution which has high requirements for agent used to implement it.

The structure of this paper is as follows.

Section 2 is a brief description of the traditional VMI concept in inventory management. It also contains a proposition of how to utilize intelligent software agents (the agent-VMI concept) to familiarize this concept to a broader spectrum of companies, especially in the sector of small and medium enterprises (SME).

In section 3, the MDV agent model is presented. It is based on the concept of multi-dimensional versioning and code segmentation. The roles of subsequent agent parts are described in detail. Moreover, the process of agent migration, environmental adaptation, agent knowledge and functionality management as well as code discarding are discussed.

The fourth section deals with an functionality algorithm of the proposed agent model and describes the whole procedure which takes place when software agents manage the inventory levels in supply chains. It contains a step-by-step description of the proposed approach.

Section 5 concludes the paper, emphasizing the advantages of utilizing intelligent software agents as the communication platform in the VMI model, especially those designed according to the MDV model.

2 Inventory Management Issues

Vendor Managed Inventory (VMI) is a concept in inventory management or a supply chain practice where the vendor manages the inventory on behalf of the consuming organization [2, 3, 4]. Normally when a distributor needs a product, he places an order against the manufacturer. This gives him total control of the timing and quantities of subsequent orders being placed [5]. In VMI, it is assumed that the vendor has much better knowledge and experience, concerning the demand for a given product, therefore he is able to manage the needed stock levels of his customers more effectively. However to be able to do so, a constant data flow, between the customer and the supplier, is essential. It has to contain information about the current inventory and sales levels of the vendors clients [6, 7]. Usually the vendor receives all necessary electronic data via EDI or the internet which he imports into his electronic information system – the partners however must agree at the outset on order quantities, order points, rules of replenishment and inventory turns [8]. He then generates the orders, based on the given data therefore scheduling his production plan to meet the required quantity. The whole business process fed by this data is relatively simple. The supplier reviews the information that has been sent in by the customer to determine if an order is needed. This review of the data varies by supplier and the software being used, but, many things are consistent.

- The first step is to verify the data as accurate and meaningful. Depending on the software, much of this verification is automated.
- On a scheduled basis, the software calculates a reorder point for each item based on the movement data and any overrides contributed by the customer or supplier. These overrides might include information such as promotions, projects, seasonality, new items and so forth.
- The VMI software compares the quantity available at the customer with the reorder point for each item at each location. This determines if an order is needed.
- The order quantities are then calculated. Typically calculation of order quantities takes into account such issues as carton quantities and transaction costs [9].

This procedure is based on the simple concept of minimal inventory levels and does not take into account any random events such as changing demand, weather conditions (for example for transportation purposes) etc.

The benefits of successfully implemented VMI which can occur are the following:

- lower customer inventories – are caused by a better organization of inventory management due to limited inventory item controlled by the vendor;
- lower supplier inventory – under VMI, the vendor knows the exact demand which he must meet, therefore he can easily lower his inventory levels to minimum;
- lower stock levels through-out the whole supply chain [10, 11] – a key issue in supply chain management is to satisfy the customer needs. Traditionally this issue has been solved by maintaining a high stock level to meet any demand fluctuation. Taking the 2 arguments stated above, VMI enables to lower the stock levels of certain entities in the supply chain, thus lowering the overall inventory costs of the whole supply chain.
- lower administrative costs – this benefit occurs to both parties engaged in VMI, and is caused by the usage of automated systems instead of regular employees.
- increased sales – those are caused by the fact, that the inventory levels are set to an optimal level, therefore decreasing chances of stock-out situations.

There are however some issues that must be taken into consideration when planning to implement the described concept. In the business environment, VMI is generally implemented by companies which satisfy the following two conditions:

- The company is a strong player, big enough to force the usage of VMI in their supply chain, therefore extorting its usage on their suppliers or clients (depending on the company's profile). It leads to situations where business partners have to adjust their IT systems to meet VMI requirements.
- The enterprise has enough free resources to invest in perfectly working electronic communication systems, which enable an undisturbed and constant information flow within the supply chain which is generally essential for VMI to work.

Moreover to make VMI work, the parties have to agree on the following terms:

- Clarify expectations - There needs to be thorough discussion about how the system will benefit both organizations in the long term.
- Agree on how to share information – if the supplier and customer can agree to share information vital to restocking in a timely manner, then the odds of a synchronized system will dramatically improve.
- Keep communication channels open – When the two parties set out to implement a VMI program, they need to meet and discuss their goals and how they need to proceed in order to realize those goals [12].

The conditions above restrain the potential usage of VMI to large enterprises, thus limiting the potential benefits of the described concept. One of many ways of enabling a brighter implementation possibility for firms, would be the application of intelligent software agents as the base of communication flow between partners (instead of traditional channels). It would also lead to reductions of the initial costs of VMI implementation, therefore giving more companies (especially in the SME sector) a chance to benefit from the concept. The proposed solution – the *Agent Managed Inventory* (AMI) model is a modified version of VMI, in which the responsibility for the re-order process still lies on the vendor, however the negotiations and calculations are performed by software agents which take actions on his behalf. Because of the fact, that each customer can posses different types of information systems for storing operational data, the used agents would have to be prepared for each possible variant of data. Therefore making them large programs, not necessarily mobile. Because of this fact, the agents used in VMI should be designed according to the MDV model described in the next chapter. This approach makes them flexible enough to handle any data structure they could encounter while still ensuring fast agent transfers. The whole implementation model has been presented in detail in chapter 4.

3 The Multi-Dimensionally Versioned Software Agent

According to the MDV agent model, an agent is composed of a sort of agent head, called *bootstrap agent*, and an agent body. The bootstrap agent is relatively small, thus it can be highly mobile. Its goal is to move over the network and to decide whether a newly visited environment is potentially interested taking into account the agent mission defined by the user. If it is, then the bootstrap agent recognizes the specificity of the environment, afterwards it communicates with a so called *proxy agent*, residing on the origin computer (i.e. a computer in which the agent has been created by the user), asking it to send an appropriate part of the code of agent body to the bootstrap agent. The bootstrap agent communicates directly with the proxy agent using messages in an agent-communication-language (ACL) [13, 14].

It may happen that after some time, if the results of agent (i.e. bootstrap agent extended by agent body) activity are satisfactory, the bootstrap agent again communicates with the proxy agent asking for the next part of agent body. This situation will be explained later.

If the agent's mission in the currently visited environment is completed then the agent body is removed from the environment and the bootstrap agent either migrates to a new environment or it returns to the origin computer in order to merge with the proxy agent.

For the sake of platform independence and security, we assume that the bootstrap agent is interpreted by the visited environment. In other words the bootstrap agent is a source code, rather than binary code.

There are four basic functions provided by the stationary proxy agent:

- It serves as a communication channel between the user and bootstrap agent: it knows the current location of the bootstrap agent and can influence the path of its movement, as well as tasks performed by the bootstrap agent.
- It encompasses all variants of the agent's code that model the variability of agent behavior. Depending on what environment is currently visited by the bootstrap agent, and according to the bootstrap agent's demands, the proxy agent sends a relevant variant of agent code directly to the bootstrap agent, thus enriching it with the skills required in this new environment and, as a consequence, enabling it to continue the global agent mission. Notice, that code transmission redundancy is avoided, since the unnecessary code (not relevant agent variants) remains together with the stationary component.
- It assembles data items that are sent to it directly from the bootstrap agent, extended by a proper agent variant, which are not useful for mobile code, however, it could be interesting to the user. The data assembled is stored in a so called *knowledge repository*.
- Whenever required, the proxy agent responds to the user who can ask about mission results and data already collected (e.g. a percentage of data initially required). If the user is satisfied with the amount of data already available, proxy agent presents the data to the user and finishes its execution (together with the mobile component).

To summarize the aforementioned discussion, one can easily determine the behavior and functions of a moving component (code). When it migrates to the next network node, only the bootstrap agent is transmitted, while the agent variant is just automatically removed from the previous node. There is no need to carry it together with the bootstrap agent, since it is highly probable that a new environment requires a different agent variant. When migration ends, the bootstrap agent checks its specificity, and sends a request for a corresponding agent variant transmission, directly to the proxy agent. When the code is completed, the mobile agent component restarts its execution.

During the inspection of consecutive network nodes only the information that enriches the intelligence of a moving agent is integrated with the agent bootstrap (thus it can slightly grow over time). Pieces of information that do not increase the agent intelligence are sent by the mobile component of the agent directly to the stationary part, in order to be stored in the data repository managed by it. This agents feature, namely getting rid of unnecessary data, is called self-slimming.

Now we focus on possibilities of the MDV`s agent versioning, i.e. on the content of the proxy agent that is always ready to select a relevant piece of agent code according to the demand of the bootstrap agent. We distinguish three orthogonal dimensions of agent versioning:

- agent segmentation,
- environmental versioning,
- platform versioning.

Agent Segmentation. Typically an agent mission can be achieved by performing a sequence of relatively autonomous tasks (or stages). Thus the agent code is divided into so called segments corresponding to consecutive tasks that have to be realized. If one task is finished successfully, then the next one can be initiated. Thus, next segment of agent code is received from the proxy agent and agent execution switches to this new segment. Depending on whether agent behavior is sequential or iterative, the previous segment is automatically deleted (in the former case) or it remains in the execution environment (in the latter case).

For example, the first segment of the agent can be used for browsing offers available at e-marketplaces. If there is an interesting offer, then second segment is transmitted on demand which is responsible for negotiation of terms of cooperation with the agent presenting the offer. And again, if the negotiation succeeds, then third segment is transmitted which is responsible for business contract signing, and so on. Contrarily, if a particular agent task fails, there is no need for the transmission of subsequent agent segments.

In the above example a sequential behavior of the agent would be recommended, in order to reduce segment transmitting. It means that the first segment should browse all offers available at the e-marketplace, mark potentially interesting offers (i.e. worth to negotiate), and then switch to the second segment. Similarly the negotiation should be performed with all agents related to the marked offers. As before, if negotiation succeeds, the second agent segment marks the respective agents, which are necessary for the execution of the third segment.

To summarize, this versioning dimension, namely segmentation, models multi-stage nature of MDV agents.

Environmental Versioning. The bootstrap agent can visit different environments providing different services, e.g. e-marketplaces, e-auctions. Moreover, every service can be implemented in a different way. For example, the e-marketplace can be implemented as a web-site, off-line database or specialized communicators. Thus, depending on the specificity of environment being visited different versions of agent segments are required.

This versioning dimension, namely environmental, models polymorphic nature of MDV agents.

Platform Versioning. Finally, every version of every agent segment is available in potentially many variants which are implemented for a particular target environment (i.e. for a particular hardware which runs agent environment: processor, operating system, network communication protocols etc.). There is one particular variant for each agent segment version, which is in a source form. It is sent to the environments which for the security reasons do not accept binary (executable) code, it means which interpret agents instead of running compiled code. Besides this particular variant, the proxy agent keeps potentially unlimited number of binary variants. If the environment accepts binary code than the proxy agent delivers a variant matching hardware and system software parameters of this environment.

Of course, binary variants ensure efficiency of the agent execution, while a single source variant guarantees security of agent interpretation.

To summarize, this versioning dimension, namely platform, models platform independent nature of MDV agents.

Now let us illustrate the behavior of a MDV agent by the example which shows step-by-step typical scenario of migration of code and data over the network.

1. The user creates a MDV agent on his/her computer using a relevant application, defines its mission, environments to be visited, assigns his/her certificate, and finally disconnects or shuts down the application.
2. The MDV agent starts its execution on the origin computer. It is composed of the bootstrap agent, the proxy agent and the knowledge repository.
3. The bootstrap agent migrates through the network to the first environment under investigation.
4. The bootstrap agent is interpreted by the visited environment; it checks its specificity and platform details; assume that it is e-marketplace implemented as a web-site, running at the computer equipped with Intel Core i7 Processor managed by MS-Windows 7 operating system; the environment allows for binary agents execution.
5. The bootstrap agent contacts the proxy agent through the network asking for the first agent segment to be send, in a version relevant to the e-marketplace being visited, and in a variant matching the hardware and software parameters already recognized.
6. First segment of the agent migrates to the environment visited and starts its execution – let's say offer browsing. The data collected by the segment which are not required in further execution, e.g. offers that could be interesting in the future, are sent back directly to the knowledge repository, thus self-slimming the agent. All offers that should be negotiated are marked by the agent segment which finishes its execution (its code is deleted).
7. The bootstrap agent contacts the proxy agent again asking for the second agent segment to be send, for example, a segment responsible for negotiation.
8. The second segment arrives and starts its execution. For the sake of simplicity we assume that negotiation fails with all respective agents related to the offers marked by the previous segment.
9. The bootstrap agent is informed about negotiation results and the second segment is deleted.
10. The bootstrap agent migrates through the network to the second environment for further investigation. Steps 4 – 10 are repeated in this environment.

4 Implementation Model

The Vendor Managed Inventory concept implemented using intelligent software agents works according to a simple procedure. To clarify the proceedings, the whole procedure had been illustrated below in the form of pseudo-code. First the vendor physically performs some preliminary negotiations with his business partner, to gain permission to use software agents which will work on his behalf. After that the basic procedure of inventory management can start. The vendor sends his bootstrap agent to check the inventory levels of each business partner, who agreed to enable agent-based negotiations. When the bootstrap agent arrives at the first client's server it checks for the existence of software representatives (agents) of the client, with whom he can establish a negotiation process. There are two possible outcomes of the test carried out – the agent either finds an agent (as illustrated on lines 11-27) or not (lines 29-47).

In the first case the negotiation process can start, therefore the bootstrap agent requests an appropriate part of agent code which is responsible for negotiations, from his server. He then asks the client's agent for the current inventory levels and matches them with the data he possesses about the client (the required inventory levels, etc.). If the inventory levels are optimal – the bootstrap agent deletes its downloaded parts and migrates to another location (lines 24-27). In other case it requests another segment, this time responsible for external factor analysis (for example it could have the possibility to check some weather forecasts for the next days, and calculate the impact of this factor on the whole re-order process). At this moment, the agent is ready to begin its negotiations with the client, while taking into account various factors like for example transportation delays resulting from bad weather conditions. The vendor agent offers a re-order based on previously defined conditions corrected by external factors. If the client's agent agrees to the suggested terms, an order is send to the vendors headquarters. Otherwise negotiations continue (while loop on 16-22) until the client's agent accepts the offer or totally rejects the conditions (in which case both parties – the vendor and the client are informed about negotiation failure). In case of a total reject, the vendor agents performs its slimming procedure and migrates to another client.

In the second case, where the agent wasn't able to find any other agents to negotiate with, it searches for ready-made data which he can analyze. This data can be in various form, it can be for example an xml file available to the agent. If no data is provided the agent informs both parties about analysis failure and migrates. If the agent finds an appropriate piece of data, it requests subsequent agent parts responsible for data analysis. He then checks the inventory levels of his client and matches them to previously defined conditions. If a re-order is needed, the agent checks if it has permissions to make a decision on his own, or if he needs human approval from either party (the vendor or the client). If the agent possesses all necessary permissions it decides how much and when to order, in other case he prepares a ready order and sends it for approval to his supervisor (the party responsible for accepting or rejecting the offer).

To clarify the described procedure, it has been presented below in the form of pseudo-code representing the overall algorithm, according to which the vendor agent takes his actions.

```
1 BEGIN
2  DEFINE FUNCTION place_order_procedure()
3   conditions = check_prerequisites()
4    #check weather, transport availability etc.
5   place_order(conditions)
6   delete_agent_body()
7   migrate()
```

```
IF(environment IS handlable)
  request_appropriate_agent_version()
  IF(agent_to_negotiate_exists())
   request negotiation module for customer
   check customer inventory levels
   IF(inventory below customer norms)

    WHILE(negotiate) #negotiation loop
     IF(agreement has been achieved)
       place_order_procedure()
      END
     ELSE
      change_offer()
      CONTINUE

   ELSE #inventory levels optimal
    delete_agent_body()
    migrate()
    END

  ELSE #no agent found
   request data analyzing format for enterprise data
   request_profile_module_for(customer)
   IF(inventory below customer norms)
    IF(human authorization IS needed)
     send_acceptance_request()
     WHILE(wait_for_response)
      IF(request_accepted)
       place_order_procedure()
       END
      ELSE
       CONTINUE
    ELSE #no human acceptance needed
     place_order_procedure()
     END
   ELSE #no order needed
    delete_agent_body()
    migrate()
    END

 ELSE
  migrate()
  END
END
```

The procedure illustrated above concludes the presentation of the AMI model.

5 Conclusions and Future Work

The proposed inventory management methodology based on software agent characterizes itself with some subsequent pros in comparison to its traditional equivalent. Those advantages are the following:

- It bypasses the standard issue of software integration between business partners (which is related to differences between both – software and hardware used by business partners)
- It automates the whole process of re-orders if the agents gain sufficient decision making permissions.
- It takes into account external data like weather conditions to improve the accuracy and quality of placed orders.

The main goal of the proposed model was to show the possibilities of using intelligent software agents to improve the efficiency of inventory management in enterprises, which not necessarily have the desire to cooperate with business partners in a traditional way or simply don’t own enough money to invest in software integration. Therefore the AMI model can be viewed as the next generation of software solutions in the area of inventory management in supply chains. Making the concept of traditional VMI available to more and more enterprises.

The proposed model is currently being modeled and tested within the NetLogo programmable simulation environment. Future work will therefore concentrate on the final implementation of the described model and an architecture proposition for real-life usage of the AMI.

References

1. University of Puerto Rico, http://repositorio.upr.edu:8080/jspui/
2. 12Manage, http://www.12manage.com/methods_vendor_managed_inventory.html
3. Skjott-Larsen, T., Jespersen, B.D.: Supply Chain Management. In Theory and Practice, Copenhagen (2005)
4. Tempelmeier, H.: Inventory management in supply networks: problems, models, solutions, Nordestedt (2006)
5. Vendor Managed Inventory, http://vendormanagedinventory.com/
6. Diedrichs, M.: Collaborative Planning, Forecasting, and Replenishment (CPFR), Muenchen (2009)
7. Li, L.: Supply Chain Management: Concepts, Techniques and Practices Enhancing Value Through Collaboration, Singapore (2007)
8. Gourdin, K.: Global Logistics Management, Oxford (2006)
9. Datalliance consulting, http://www.datalliance.com/vmi.pdf
10. Waller, M., Johnson, M.E., Davis, T.: Vendor-Managed Inventory in the Retail Supply Chain. Journal of Business Logistics 20, 183–203 (1999)
11. Enarsson, L.: Future Logistics Challenges, Copenhagen (2006)
12. NC State University, http://scm.ncsu.edu/public/lessons/less030305.html
13. FIPA, http://www.fipa.org/specs/fipaSC00001L/
14. FIPA, http://www.fipa.org/specs/fipa00061/

An Improved Data Warehouse Model for RFID Data in Supply Chain

Sima Khashkhashi Moghaddam, Gholamreza Nakhaeizadeh,
and Elham Naghizade Kakhki

Department of Computer Science, Azad University of Mashhad, Iran
Karlsruhe Institute of Technology, Germany
Department of IT Management, Shahid Beheshti University, Iran
{Simamoghaddam1983,e.naghizadeh}@gmail.com,
nakhaeizadeh@statistik.uni-karlsruhe.de

Abstract. Nowadays, one of the fundamental challenges which exists in applying RFID technology is optimal managing of a large amount of data which are produced and gathered to exploit RFID. The notion leads to the decrease of information systems productivity and also reduces the information efficiency. Recently, some efforts have been made in order to introduce different data warehouse models for RFID to utilize information. Our research aims at introducing an improved model for RFID data warehouse in supply chain. The research has manifested a new framework for managing a large amount of RFID data. Firstly, an effective model of coding data path in supply chain is initiated and then in order to retrieve time according to coding path, a numerical model of XML environment has been used. Finally, by utilizing an index technique, RFID data is aggregated which caused a considerable reduction in the volume of data and also made a dramatic fall in the response time of queries on RFID data.

Keywords: Data Warehouse, RFID, Supply Chain.

1 Introduction

In the past few years, automatic identification technique has been applied in supply chain more than before. Factually, the technology provides the possibility of tracing, gathering and managing information of items which are moved through supply chain. Radio frequency identification (RFID) has mostly grown in automatic business area. One of the typical ways of saving information in data warehouse is exploiting dimensional method according to data cube and fact table but the method cannot be used for RFID data in supply chain. In order to make it more clearly, it is assumed that we have a fact table with the following dimensions.

```
(EPC, location, time_in, time_out: measure)
```

Data cube calculates all possible groups in fact table by gathering the records with same value amongst all possible combinations of dimensions. For instance, we can

J.-S. Pan, S.-M. Chen, N.T. Nguyen (Eds.): ACIIDS 2012, Part I, LNAI 7196, pp. 488–497, 2012.

achieve the number of goods which has been in a particular location at the same time, but ignoring the relationship among records is the problem of the model. So, it is difficult to find the number of P goods which was sent from L distribution center to V store. There have been some P goods everywhere but we do not know how many of them have been sent to location V. This urges to the need for a stronger path-driven model for data aggregation. The main challenge of this research is the optimal management of a large volume of generated data through RFID application which sometimes reaches to a couple of terabytes per day. Therefore, the method of saving and categorizing data is not primarily important; in fact the most crucial and difficult issue is supporting high-level queries in a similar as well as a productive way. Some of the queries require a complete review of imported RFID database. In order to achieve such goal, the researchers propose a framework for managing and saving data based on compressing method which results in efficient processing of RFID data queries. The model provides the possibility to keep a path for each item by utilizing path coding methods in XML tree as well as prime numbers attributes. The model not only uses bulky movement of items through a supply chain but also provides the possibility of data aggregation according to other criteria apart from their movements in supply chain. Furthermore, a new method for incremental aggregation of data according to various combinations of RFID data queries is proposed which analyses data based on different dimensions in addition to the path dimension. Finally, the structure of tables in data warehouse and the last architecture of the proposed model are provided.

2 A Review of Previous Studies

Most topics in the field are focused on data model definitions and compressing methods which guarantee the effective query processing of RFID data. These researches are divided to two main categories: While the first category concentrates on on-line processing of RFID data and data stream processing [4] [11], the second group focuses on offline details and effective methods of saving data [3] [1] [5]. Most of these researches have assumed that in supply chain, goods start moving in large groups at first and then through the paths, they divide into smaller groups. Based on this idea, data compressing depends only on nodes of supply chain and movement of objects usually is not considered. Consequently, if any change occurs in supply chain scenario or movement form, the compressing way will not be effective anymore and the volume of tables will not be well-reduced. Besides, in this data aggregation method, data increase leads to an extra cost regarding joining tables. In addition, in those cases that performing various levels of granularity and complexity is required, most of these approaches show less flexibility with regard to dimensional queries. A new storing model has been suggested by Gonzalez [5] in order to provide effective support for path oriented aggregate queries. In his model, a table called “stay” is presented which is structured as the following:

```
Stay - table (GID, Loc, Start-time, End-time, Count)
```

The table is intended to store RFID data in an effective manner. In most RFID Applications, goods usually move in large groups at earlier moving levels of supply chain. These groups divide into smaller ones later in the path. Consequently, each row of stay table manifests goods which have moved from a similar location simultaneously. According to this method, GID is an indication of smaller GIDs or a list of EPCS. As a result of applying this method, the volume of data can be decreased considerably. But if goods do not move in large groups, the volume of basic table does not decrease. Besides, in the model, GID is designed to point at a list of fields which have string format. In other words, each GID includes some GIDs and EPCs which are strings. Hence in order to join tables, string assessment is needed which requires so much time compared to numerical case. The other shortcoming of this model is that it takes longer to process queries of the first group, i.e. tracking queries. Because, in the method the location of each group of goods (GID) in stay table is kept, in order to find movement path a table needs a sequent joining table of stay and map tables. One of the other important problems which can be observed in most of the models presented for RFID data is their low flexibility in comparison with the queries which are presented in various data dimensions and sub-dimensions. Moreover, these models are not able to change the levels of data granularity. In other words, we need a model which not only categorize data according to common movement of items but is also able to categorize data according to various combinations of its attributes.

3 Categorizing Queries in RFID Systems

In 2008, Lee and Chung worked on categorizing the queries on RFID data in their article [7]. From then, the response time to the queries has been chosen as a criterion for comparing different data models which are being proposed for RFID data. In fact they suggested a query framework for RFID data which encompasses two query groups.

1. Tracking Queries: presents the background of the movement of a tag.
2. Path – oriented Queries: which is divided into two parts:

- Path – oriented retrieval Queries: they should find the tag with specific conditions (for instance, time or location conditions)
- Path – oriented Aggregate Queries: presents the aggregate value of tags which meet specific conditions.

4 Proposed Model

4.1 Coding the Path for the Movement of RFID Tag

In their article [11], Wu, Lee and Hsu has applied the uniqueness of prime numbers as a means of showing the relationship between existing elements in a XML tree, which, in our model is utilized for coding nodes of the path tree of RFID tags. Reference [6] provides more information regarding the application of prime numbers in the coding nodes of the path tree.

There are various methods for coding the path [7] [8] [9] but compared to [11], our applied method provides a more effective way of processing to RFID queries. Moreover, concerning the large amount of RFID data, we have tried to apply a method which firstly does not require storing a lot of information in order to retrieve the path and secondly, retrieving the path is possible by running simple queries without recurring joins of path table [6].

Path information can be kept according to the path coding, but saving the related time for each location has not gained yet. In order to reach that, another tree is created which is called time tree in which each node keeps location, time-in and time-out of that location. In time tree two nodes are considered similar when they have a similar path (similar location) sequence to the node) as well as same time-in and time-out. The way of building this tree is to some extend similar to the way of building the path tree; however, those nodes with the same location but different time-in and time-out information are considered diverse. For efficient retrieving of time information, values of each node should be accounted by using Region Number Scheme [12].

In order to keep path and time information of tags, two tables, path- table and time-table, are created according to the above-mentioned methods; path-table keeps path information and time-table saves information of the time tree. These tables are structured as follows:

- Path- table: Path ID, Element-Enc, order-Enc
- Time-table: start, End, location, start-time, End-time.

4.2 Data Aggregation

One of the most crucial problems in most of the data models which have been proposed for RFID data is the low flexibility of them in terms of with queries which are provided in different combinations of data attributes and dimensions. In other words, these models are not able to change the levels of granularity; as a result, the response time to the aggregate queries escalates. In order to support this group of queries, a model is required which has the ability of compressing data not only based on path or time elements, but also on each and every attribute, as needed. In order to obtain this flexibility in data aggregation, the concept of aggregation factors has been used [4]. To understand the concept better, take location dimension for instance; Data can be aggregated on city, region and country levels. Also goods can be categorized according to its brand, type and price. Imagine a set like S which contains all the attributes of the input data:

$$S= \{A1, A2 \ldots An\}$$

In other words, all attributes are groups in S set according to various dimensions such as location, time and items. In addition, for each dimension, a relation is defined as following:

R= {(x, r, y): x, yes}

r indicates a semantic relation between attributes such as part-of or is-a. By using this method, some taxonomies can be made from main attributes. Fig. 1 illustrates one of these categories in which thicker vectors are signs of is-a and thinner vectors are indicators of part-of relation.

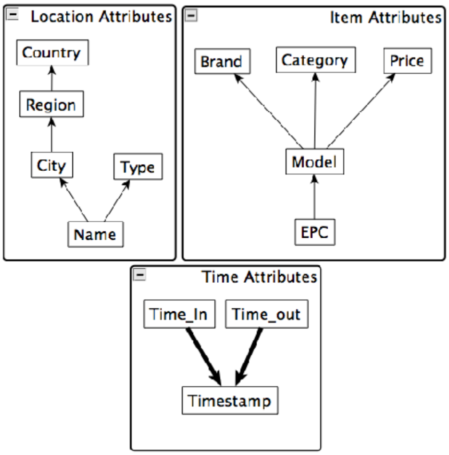


Fig. 1. An example of taxonomy

Aggregation factor which is shown by Q encompasses a set of A1, A2 …An attributes. Q depends on the categories which are defined based on different dimensions.

Concerning the above-mentioned definition of aggregation factor and its application in creating aggregated record the following four tables are required for saving information:

1. Attribute- table: AttrID, Name.
2. AttributeList-table: AttrlistID, AttrID, Value.
3. Item – table: EPC, stockID, AttrlistID.
4. Location- table: locationID, Name,AttrlistID

4.3 The Final Aggregate Data Warehouse Tables

By combining the result table of the first part of our proposed model (path coding) and second part (data aggregation) the final structure can be gained as below (Fig. 2). In this structure which includes seven tables, all of the fields and their relations are determined.

It should be noted that stock table is the result of applying aggregation factor on the middle table which is called stay and will be discussed in more details later.

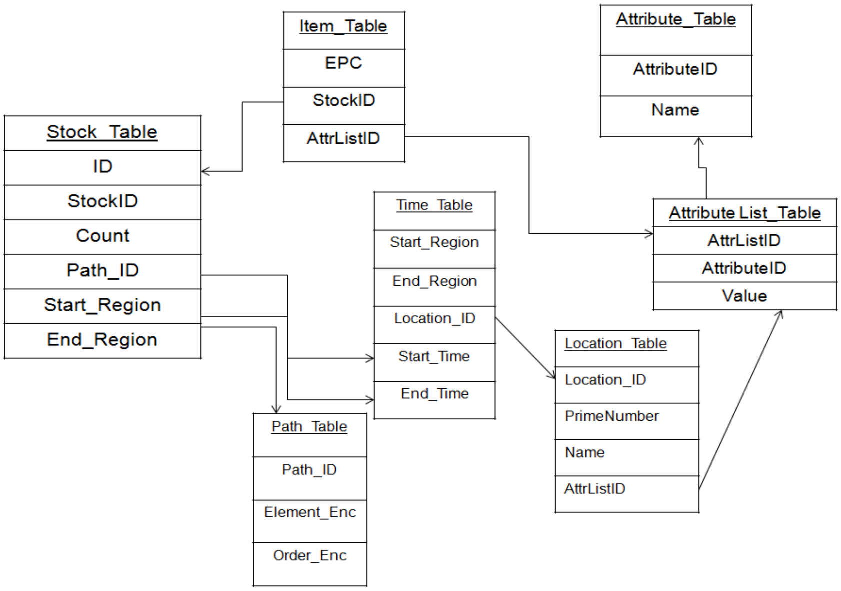


Fig. 2. Storage schema

4.4 Final Architecture of RFID Data Warehouse Model

In Fig. 3, the final framework of proposed model is shown. The structure illustrates the process of data organization which includes the following six steps:

1. The implemented system loads an XML format file including supply chain path structure. Then, the application creates supply chain path tree according to this structure and fill the Path table.
2. In this part, through the XML file, the user should feed system with a set of existing attributes along with their values. Attribute and AttributeList tables are filled through the XML file. The user should choose the aggregation factor according to the input attributes.
3. The system receives raw data. The data has some problems such as data redundancy and repetition. Hence, in this step data is cleaned and ordered according to time-in. The result of this step is RFID Data table with the following records: RFID Data: EPC, Location, Time-in, Time-out, Attribute1, Attribute2… AttributeN.
4. RFID Data table is fed and through running location and time coding algorithm, time tree is created and finally the Time table is filled. Data of this step is kept in stay table which has the following structure: Stay- table: EPC, Path ID, Start-region, End-region, Attribute1 …AttributeN.
5. The application aggregates Stay table data by applying aggregate factors which were entered by user in the second step. Consequently, Stock table is filled.
6. This architecture, finally, provides the possibility of running some the important RFID queries through a proper user interface.

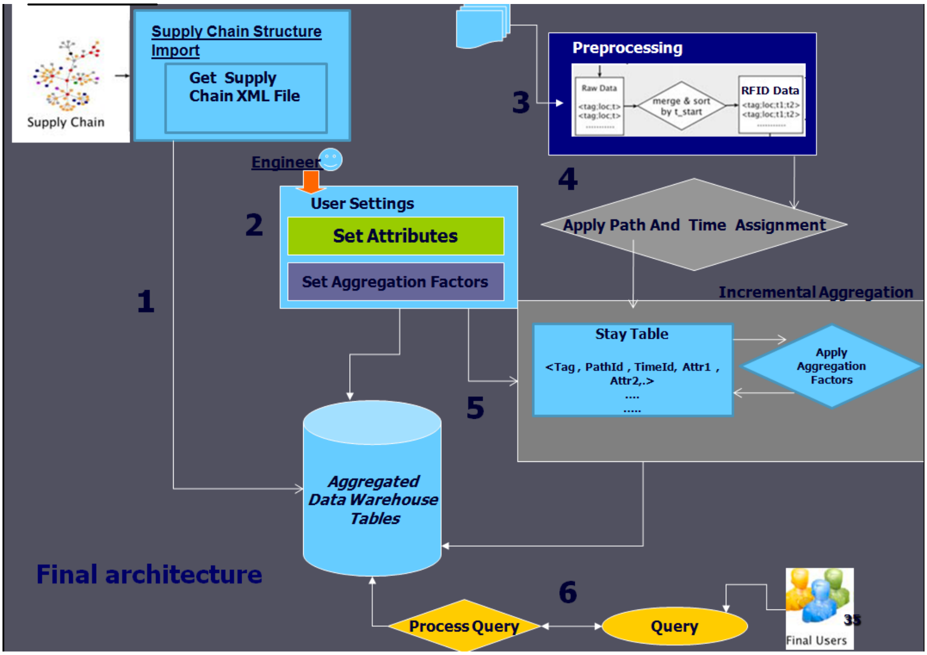


Fig. 3. Final architecture

5 Running the Model and Its Comparison with Other Alternatives

In order to implement the model Visual Studio.Net environment and C# programming language is exploited. SQL server 2008 program is used for table creation and our proposed model is compared with Gonzalez model.

Five databases, RFID 200, RFID400, RFID600, RFID800, and RFID1000, were created and filled by RFID Data[1] for testing the model. Each of these databases has two other tables which are called Stay and RFID data in addition to the aforementioned tables (4.3). 200000, 400000, 600000, 800000 and 1000000 RFID cleaned records were put in each database respectively. Then RFID Data Manager Application was run on each of the abovementioned databases alternatively until all of their tables were filled according to the aggregate factor of:

Q: {path and time-in, time-out and model}.

In the following step, three query categories are defined based on [6] in order to examine and compare the response time of these databases to important RFID queries. Among these three categories, the first one includes tracking Queries (chart. 1), the

[1] Professor Roberto De Virgilio from Roma TreUniversity, kindly provided us with 1 million RFID records which described the movement of goods through a supply chain.

second one includes path-oriented retrieved queries (chart. 2, chart. 4) and the third one includes path-oriented aggregation queries (chart. 3). The fourth chart is location-oriented aggregation queries.

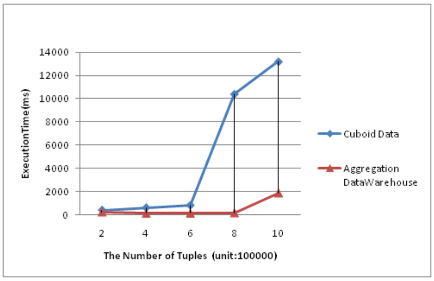


Chart 1. A Sample of tracking query comparison

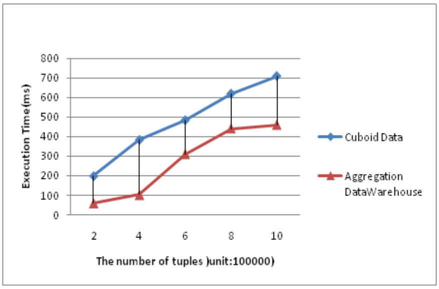


Chart 2. A Sample of path oriented retrieval query comparison

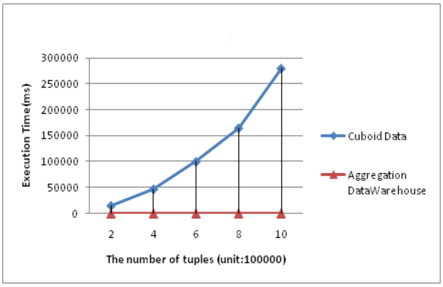


Chart 3. A Sample of path oriented aggregate query (path condition) comparison

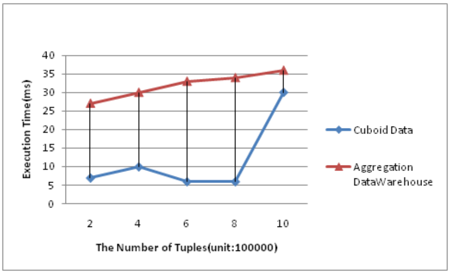


Chart 4. A Sample of path oriented aggregate query (location condition) comparison

6 Conclusions

As shown in the charts 1, 2 and 3, the proposed model performs considerably better than Gonzalez dimensional model regarding tracking queries, path-oriented retrieval queries and path-oriented aggregation queries. But on the other hand, Gonzalez dimensional model has a better performance concerning location-oriented queries than our proposed model.

Generally, dimensional data warehouse -which RFID-Cuboids model is a kind of them-returns a proper time in SQL aggregation queries. But in comparison with our proposed model, RFID-Cuboids model performance is only better in those aggregation queries which merely imply location conditions.

7 Further Suggestions

Utilizing more advanced lossless compression techniques in order to reduce the required storage is suggested. RFID data requires an efficient framework in order to answer high-level queries. Especially we are interested in dynamic scenarios where supply chain can change in terms of object transitions and topology. It needs a flexible and dynamic aggregation mechanism, and also a run-time query engine which is able to update tracking information. In this paper we proposed a data warehouse model for RFID data, which is applicable to other types of sensor data or spatial data, therefore, this aspect needs to be investigated either.

References

[1] Agrawal, R., Cheung, A., Schonauer, S.: Toward traceability across sovereign distributed RFID databases. In: 10th International Database Engineering and Application Symposium, IDEAS 2006, Delhi, India (2006)

[2] Bai, Y., Wang, F., Liu, P., Zaniolo, C., Liu, S.: RFID data processing with a data stream query language. In: The 23rd International Conference on Data Engineering, ICDE 2007, Istanbul, Turkey (2007)

[3] Ban, C., Hong, B.-H., Kim, D.: Time Parameterized Interval R-Tree for Tracing Tags in RFID Systems. In: Andersen, K.V., Debenham, J., Wagner, R. (eds.) DEXA 2005. LNCS, vol. 3588, pp. 503–513. Springer, Heidelberg (2005)

[4] De Virgilio, R., Sugamiele, P., Torlone, R.: Incremental Aggregation of RFID Data. ACM, IDEAS, Calabria, Italy (2009)

[5] Gonzalez, H., Han, J., Li, H., Klabjan, D.: Warehousing and Analyzing Massive RFID Data Sets. In: The 22nd International Conference on Data Engineering, ICDE 2006, Atlanta, USA (2006)

[6] Moghaddam, S.K.: An Improved Data Warehousing Model for RFID Data of Supply Chain. Master thesis. Islamic Azad University Of Mashhad, Iran (2011)

[7] Lee, C., Chung, C.: Efficient storage scheme and query processing for supply chain management using RFID. In: The 34th International Conference on Management of Data, SIGMOD 2008, Vancouver, Canada (2008)

[8] Min, J., Park, M., Chung, J.: A queriable compression for xml data. In: SIGMOD (2003)

[9] Rao, P., Moon, M.: Indexing and querying xml using prufer sequences. In: ICD (2004)

[10] Wang, F., Liu, P.: Temporal management of RFID data. In: VLDB 2005, pp. 1128–1139 (2005)

[11] Wu, X., Lee, M., Hsu, W.: A prime number tagging scheme for dynamic ordered xml trees. In: ICDE (2004)

[12] Zhang, C., Naughton, J., DeWitt, D., Luo, Q., Lohman, G.: On supporting containment queries in relational database management systems. In: SIGMOD (2001)

Author Index