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

LNAI 7198

Intelligent Information and Database Systems

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

3 Part III

Springer

Lecture Notes in Artificial Intelligence 7198

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 III

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 978-3-642-28492-2 e-ISBN 978-3-642-28493-9
DOI 10.1007/978-3-642-28493-9
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 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

The Impact of Human Brand Image Appeal on Visual Attention and Purchase Intentions at an E-commerce Website

Young Wook Seo[1], Seong Wook Chae[2], and Kun Chang Lee [3,*]

[1] Software Engineering Center at NIPA, Seoul 138-711, Republic of Korea
[2] National Information Society Agency, Republic of Korea
[3] SKK Business School and Department of Interaction Science, Sungkyunkwan University, Seoul 110-745, Republic of Korea
{seoyy123,seongwookchae,kunchanglee}@gmail.com

Abstract. The purpose of this study was to examine how human brand image appeal affects visual attention using eye-tracker, a visual attention measuring apparatus, at an e-commerce website. Additionally, we examined the effect of human brand image appeal on purchase intention using a questionnaire method. We conducted an eye-tracker experiment, collected survey data, and conducted interviews with each participant using laptop and perfume products. After collecting 108 valid data, the human brand images were divided into three groups: high human brand image appeal group, low human brand image appeal group, and no human brand group. We applied MANOVA and t-tests to analyze the data. The results showed that the level of human brand image appeal has a significant influence on a consumer's visual attention and purchase intention towards a product. Both visual attention (human brand and product AOI) and purchase intention are highest for the high image appeal group.

Keywords: Human Brand; Image Appeal; Purchase Intention; Eye-tracking.

1 Introduction

Electronic commerce (e-commerce) has emerged as a new paradigm in modern society. Increasing product sales through the product's brand or a human brand is clearly an important concern for e-commerce. In recent years, creating and promoting a human brand requires a significant amount of capital, and organizations are dedicated to managing human brands and building emotional bonds with consumers [8]. Moreover, with the advent of e-commerce, the potential of utilizing a human brand image in Internet technologies is considerable.

The present article focuses on human brand image appeal at an e-commerce website. Specific to the current investigation, human brand images on websites such as online shopping malls are examined in regards to how they might induce hedonic reactions of image appeal and perceived purchase intention for the user. In this study,

* Corresponding author.

J.-S. Pan, S.-M. Chen, N.T. Nguyen (Eds.): ACIIDS 2012, Part III, LNAI 7198, pp. 1–9, 2012.

human brand images refer to the representation of human brands in website images. The levels of human brand images identified in this paper are as follows: a human brand with higher image appeal for the user (high image appeal condition), a human brand with lower image appeal for the user (low image appeal condition), and a control condition with no human brand images (no human brand condition).

This research focuses on differences among the levels of human brand images by asking the following two questions:

(1) How do the use of high image appeal, low image appeal, and no human brand in an online shopping mall contribute to the creation of purchase intention?
(2) How does the level of human brand images used for products in an online shopping mall affect visual attention?

To gain insight on how Internet shopping mall users perceive human brand images as one element of e-commerce website design, a controlled experiment was conducted using a questionnaire and eye-tracking methodology. The eye-tracking method has recently been used to measure an individual's visual attention. Eye-tracking is a physiological technique used to sense visual attention by tracing eyesight.

2 Theoretical Background and Research Hypotheses

2.1 Human Brands and Image Appeal

When *Time* and *Forbes*, two magazines known worldwide, announce the most influential and popular 100 people every year, the world shows interest in who is on the list. This new social phenomenon indicates that people perceive well-known personas like celebrities, sports stars, and actors as human brands.

A brand can be broadly defined as a trade marketable visual or verbal piece of information that identifies and differentiates a product or service. Traditionally, brands have been related to products, services, or organizations, but today researchers argue that brands can also be human [2, 6]. A human brand refers "to any well-known persona who is the subject of marketing communication efforts" [8] and serves as an intangible asset, such as a social reputation, image, or credibility. For example, celebrity brands (e.g., Super Junior), athlete brands (e.g., Yuna Kim), and CEO brands (e.g., Steve Jobs) can be thought of as human brands. Today, celebrities are regarded not only as famous entertainers or sports stars, but also as human brands in people's minds. More recently, the concept of the human brand has received increased attention and has played a vital role in business. Companies are dedicated to managing human brands and building emotional bonds with consumers [8].

Image appeal refers to the extent to which images on the website are perceived as appropriate and aligned to user expectations, or as satisfying or interesting [1]. Human brand image appeal pertains to all images of human brands on the website, including product-specific images as well as any images of human brands that may or may not interact with the products. Image appeal goes beyond aesthetics or the attractiveness of images, as it also encapsulates the hedonic emotions elicited by viewing the images [1].

2.2 Purchase Intentions

Information systems research has focused on constructs such as trust, usefulness, enjoyment, and website quality as determinants of the online purchase intention [11]. Varying levels of website quality have been shown to influence online purchase intentions while conveying the same intrinsic product information, suggesting that website quality independently influences consumer perceptions [11]. Studies have shown that involvement with a website is positively related to attitudes toward the website, which in turn influences consumers' intention to purchase a product on the website [5]. Vakratsas and Ambler [9] asserted that advertisements must have some mental effects, such as cognitive involvement, in order to affect behavior. In their model, conative responses (such as purchase intention) represent the consequences of these mental effects [5], suggesting that associations between cognitive involvement and the purchase intention exist in the persuasion model of advertising [5].

2.3 Research Hypotheses

We proposed a research model (see Figure 1) consisting of three main constructs: human brand image appeal, visual attention, and purchase intention.

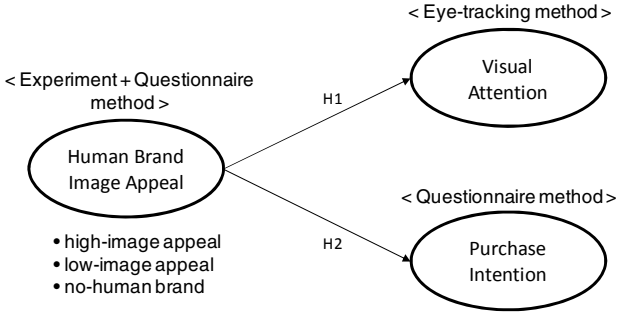


Fig. 1. Research Model

In online environments, images of people can be used to induce emotional responses, which may result in favorable attitudes toward the site [1]. Cyr et al. [1] revealed that higher human image appeal results in higher levels of trust. Wu and Lo [12] revealed that brand awareness has a significant influence on core-brand image (parent-brand image), thus indirectly affecting the core-brand attitude and impacting the consumer purchase intention towards extended products. On the basis of these studies, we propose the following hypotheses.

H1: Visual attention in on online shopping mall will be more effective in the high human brand image appeal condition as compared with the low image appeal and no human brand conditions.

H2: The purchase intention in an online shopping mall will be highest in the high human brand image appeal condition as compared with the low image appeal and no human brand conditions.

3 Experiment and Analysis

3.1 Experimental Design and Procedure

Pre-test
We implemented a pre-test using surveys and in-depth interviews with 11 participants on the types of products found in an Internet shopping mall and human brands. First, 23 human brands were recommended by 10 graduate students in order to select the human brands included on the pre-test. We then selected the 10 most favorable human brands by investigating the preferences for each one and selected two products by conducting the pre-survey, with the measurement of the product types based on a 5-point scale [10]. Based on the results of the pre-test and interviews, the laptop was selected as the functional product ($mean_{functional}$ = 4.82, $mean_{symbolic}$ = 2.24) and perfume as the symbolic product ($mean_{functional}$ = 1.68, $mean_{symbolic}$ = 4.00). In this study, Amazon (www.amazon.com) was chosen as the Internet shopping mall based on the results of the in-depth interviews.

Participants
For this experiment, 56 healthy college students at a Korean university were recruited. Prior to the experiment, the subjects were given verbal information explaining the experimental procedures. We received written informed consent from all participants and each subject was paid 10,000 Korean won for participation. Among the participants, the data were corrupted for 2 subjects, who were eliminated from the study. A total of 54 subjects (24 men and 30 women) were thus employed in this study. About 70% were aged 21 to 29 years, and 26% were aged 20 years or younger.

Materials and design
We created a number of screens with detailed product descriptions to implement the human brand experiment, simulating the Amazon shopping mall. The screen with the product explanation consists of the product, message, and human brand domain [7]. The product domain represents the product for advertising. The message domain is the text, including abstractly written sentences about the product concept along with a detailed explanation. Finally, the human brand domain uses celebrities as human brands with images in order to enhance consumer interest. Generally, consumers pay more visual attention to the product and human brand domains than the message domain. However, when consumers experience a discrepancy between existing knowledge and the product brand presented in the advertisement, they pay more attention to the message domain in order to remove the discrepancy [4]. Stimuli for the experiment are a total of 20 screen images with detailed product descriptions made by combining 2 product types (laptop, perfume) that do not have a human brand with 10 different human brands. All items in the questionnaire were constructed as agree-disagree statements on a 7-point Likert scale. The survey items we used were adapted from those found in the academic literature [1, 5, 11].

Apparatus
The online shopping mall screen shots were presented on a 19-inch monitor with a resolution of 1024 x 768 pixels. The Tobii Eye Tracker™ (X120) developed by Tobii Technology Corp. was employed to record participants' eye movements during the

experiment, and data were treated using Tobii Studio. Eye-Tracker measures participants' visual attention, which consists of eyeball fixation and saccade. Eyeball fixation shows how long a participant's eyes stay fixed at a certain area and saccade is the momentary movement between eyeball fixations. Fixations were detected at 100ms or above, which is an appropriate cut-off point for tracking eye movements in an experiment [3].

Procedures

The experimental procedure was as follows. First, before starting the experiment, we asked participants to select their favorite human brands out of the 10 provided. Second, a calibration test was conducted to trace each participant's eye movements before starting the experiment. Third, two shopping processes (functional, symbolic) were sequentially displayed to participants to simulate the real environment for online shopping. For time control of each participant's visual attention measure, each screen was presented for 10 seconds. Lastly, participants were asked to answer the questionnaire items and were then interviewed.

3.2 Experimental Results and Analysis

We used SPSS 13.0 software to analyze our measurements and test our hypotheses. We divided the human brand data into two groups based on the 7-point median value (=4.00) for image appeal after averaging the five items measuring image appeal. Consequently, the total data were divided into three groups (high human brand image appeal group, n = 43; low human brand image appeal group, n = 33; and no human brand group, n = 32). In this research, visual attention was based on the fixation length on the human brand, product, and message AOI (area of interest) in the product information display (see Figure 3(b)).

Table 1 shows the mean values of the three AOI and purchase intention. The mean fixation length of the high image appeal (HIA) group was longer than that of the low image appeal (LIA) and no human brand (NHB) groups for brand AOI and product AOI. Also, the mean value of the purchase intention was higher for the HIA group than the LIA and NHB groups (see Table 1 and Figure 2).

Table 1. Group means for the two dependent variables

Group	N	Visual Attention						Purchase Intention	
		Human Brand AOI		Product AOI		Message AOI			
		Mean	Standard Deviation	Mean	Standard Deviation	Mean	Standard Deviation	Mean	Standard Deviation
HIA	43	0.833	0.895	1.644	1.144	5.234	2.408	3.274	1.129
LIA	33	0.603	0.587	1.226	0.996	5.506	2.303	2.467	1.081
NHB	32	NA	NA	1.454	1.463	5.788	2.132	2.800	1.164

(Note1) HIA: High-Image Appeal of human brand, LIA: Low- Image Appeal of human brand, NHB: No-human brand

(Note2) Visual Attention : AOI Area Fixation Length

(Note3) NA : Not applicable

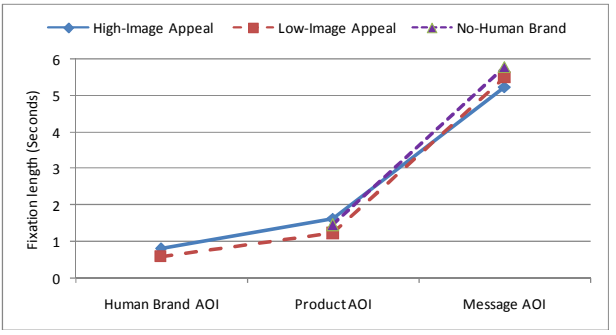
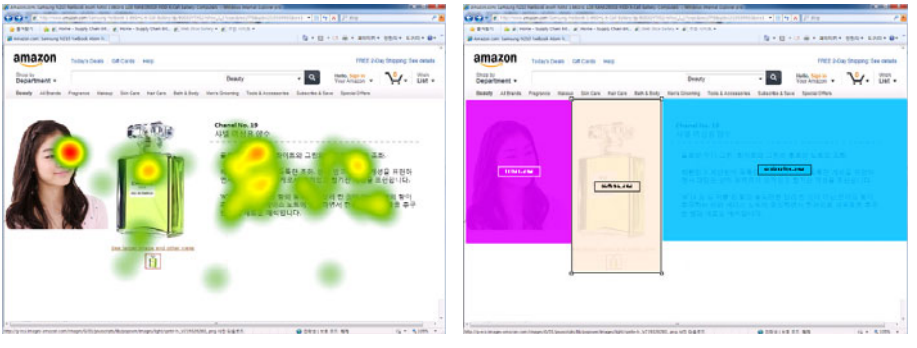


Fig. 2. Comparison of Mean-value of Fixation length for AOI

Figure 3(a) presents the heat map, which is a visualization tool provided by Tobii Eye-Tracker. It shows the average fixation length of participants, which can be interpreted as the degree of visual attention for the three areas, namely AOI. As seen in Figure 3(a), the results show that, intuitively, the message AOI received more visual attention than the other types of AOI.



(a)Heat map measured by eye-tracking (b) AOI(area of interest)

Fig. 3. An Example of Eye-tracking Result

As shown in Table 2, MANOVA was also used to examine differences between the group means for the two dependent variables of human brand images (visual attention and purchase intention). Groups were defined by the three manipulated levels of human brand image appeal (high image appeal, low image appeal, and no human brand). As shown in Table 2, the F-statistic is significant for both dependent variables, which indicates that at least one of the human brand image appeal levels is different from the others. The contrasting results shown in Table 3 support these differences. Both visual attention (human brand and product AOI) and purchase intention are highest for the high image appeal condition, with the exception of message AOI. In the high image appeal condition, a greater fixation length of human brand and product AOI is demonstrated than in the low image appeal (p<0.1) and no human brand (p<0.01) conditions. For the purchase intention, significant differences are found between the high image appeal group and both the low image appeal

(p<0.01) and no human brand (p<0.1) groups. There is no significant difference between the low image appeal and no human brand conditions.

Table 2. A Summary of the Result of the Multivariate Analysis of Variance

Dependent variable		Sum of Squares	df	Mean Square	F	Sig.
Visual Attention	Human Brand and Product AOI	20.185	2	10.093	5.027	0.008*
	Message AOI	5.647	2	2.823	0.535	0.587
Purchase Intention		12.527	2	6.263	4.947	0.009*

*p<0.01

(Note) Human brand image appeal level is the independent variable.

Table 3. MANOVA Contrast Results

Contrast		Dependent variable		
		Visual Attention		Purchase Intention
		Human Brand and Product AOI	Message AOI	
High-Image Appeal versus Low-Image Appeal	Mean Difference	0.647	-0.272	0.808
	Standard Error	0.328	0.532	0.260
	Significance	0.051*	0.610	0.002***
High-Image Appeal versus No-Human Brand	Mean Difference	1.022	-0.553	0.474
	Standard Error	0.331	0.536	0.263
	Significance	0.003***	0.305	0.074*
Low-Image Appeal versus No-Human Brand	Mean Difference	0.375	-0.281	-0.333
	Standard Error	0.352	0.570	0.279
	Significance	0.288	0.623	0.235

*p<0.1, ** p<0.05, ***p<0.01

In order to statistically verify the visual attention of the human brand group, data on fixation length for each AOI (provided by Eye-Tracker) were analyzed. Table 4 shows the comparison of means among the three AOIs and the t-test results to verify whether there are differences between high and low image appeal for the human brand. The visual fixation length at the 'product area' was significantly longer in the high image appeal group than in the lower image appeal group of human brand. All the experiment results showed that hypotheses 1 and 2 are statistically supported.

Table 4. T-test Results for Visual Attention of Human Brand Group

	Mean		Standard Deviation		t-value	Mean Difference	Sig.
	HIA	LIA	HIA	LIA			
Human Brand AOI	0.833	0.603	0.895	0.587	1.347	0.230	0.182
Product AOI	1.644	1.226	1.144	0.996	1.666	0.417	0.099*
Message AOI	5.234	5.506	2.408	2.303	-0.497	-0.272	0.620

*p<0.1

(Note1) HIA: High-Image Appeal of human brand, LIA: Low- Image Appeal of human brand

4 Discussion and Conclusion

In this study, we investigated the effects of a human brand on the visual attention and purchase intention of consumers in an online shopping mall by comparing the group with high human brand image appeal with groups having low human brand image appeal or no human brand. The greatest contribution of this study was that an eye-tracker was employed to analyze the visual attention of consumers on the screen displaying detailed product descriptions in an online shopping mall in order to overcome the limitations of a questionnaire survey.

The results of this study are summarized as follows: First, the screen on which human brand had higher image appeal was more effective in the aspect of visual attention of the consumers as compared with screens presenting lower image appeal or no human brand. In particular, the visual fixation length at the 'product area' was longer for the high human brand image appeal group as compared with that of the lower human brand image appeal group. Additionally, the high image appeal group showed a longer visual fixation length at the region of 'human brand area plus product area' as compared with that of the no human brand group. Second, the screen on which the human brand had higher image appeal had more positive effects on the purchase intention of the consumers as compared with the screens having lower image appeal or no human brand.

The implications of this study from a practical viewpoint are as follows: First, an online shopping mall planner should consider the use of human branding for product advertisements because the use of human brand advertisements in an online shopping mall may result in higher purchase intentions as compared with advertisements lacking a human brand. In particular, a human brand should be used for expensive and profitable products, considering the high cost of the human brand, rather than applying it to all products. Second, a human brand with high image appeal, that is, a human brand that is well matched with the online shopping mall product, should be employed to draw a positive response of the consumers. The purchase intention may decrease if a human brand has low image appeal even though it may be consumers' favorite human brand. Thus, considering the aspect of cost, the online shopping mall company must verify if a human brand holds high image appeal in connection with the product before utilizing it on the website. Third, the website designer should take into account the fact that the human brand with high image appeal is associated with consumers' vision remaining fixed on the product for a longer time. In other words, the visual design and view need to be differentiated so that consumers' preferences may be increased while their vision is fixed on the product in an advertisement using a human brand with high image appeal.

This study has the following limitations: First, various product categories were not taken into account because only laptop computers and perfumes were selected for the experiment as the representative products through the preliminary investigation. Further studies may thus need to increase the possibility of generalization of the research results by including variables such as the degree of involvement of the product. Second, the sample size was relatively small and the participants consisted only of university students.

Acknowledgments. This study was supported by WCU (World Class University) program through the National Research Foundation of Korea funded by the Ministry of Education, Science and Technology (Grant No. R31-2008-000-10062-0).

References

1. Cyr, D., Head, H., Larios, H., Pan, B.: Exploring human images in website design: a multi-method approach. MIS Quarterly 33(3), 539–566 (2009)
2. Hazan, C., Shaver, P.R.: Attachment as an Organizational Framework for Research on Close Relationships. Psychological Inquiry 5, 1–22 (1994)
3. Inhoff, A.W., Radach, R.: Definition and computation of oculomotor measures in the study of cognitive processes. In: Underwood, G. (ed.) Eye Guidance in Reading, Driving and Scene Perception, pp. 29–53. Elsevier, New York (1998)
4. Janiszewski, C.: The Influence of Print Advertisement Organization on Affect toward a Brand Name. Journal of Consumer Research 17(1), 53–65 (1990)
5. Jiang, Z., Chan, J., Tan, B.C.Y., Chua, W.S.: Effects of Interactivity on Website Involvement and Purchase Intention. Journal of the Association for Information Systems 11(1), 34–59 (2010)
6. Leets, L., De Becker, G., Giles, H.: FANS: Exploring Expressed Motivations of Contacting Celebrities. Journal of Language and Social Psychology 14, 102–123 (1995)
7. Rosbergen, E., Pieters, R., Wedel, M.: Visual attention to advertising: A segment-level analysis. Journal of Consumer Research 24(3), 305–314 (1997)
8. Thomson, M.: Human brands: Investigating antecedents to consumers' strong attachments to celebrities. Journal of Marketing 70, 104–119 (2006)
9. Vakratsas, D., Ambler, T.: How Advertising Works: What Do We Really Know? Journal of Marketing 63(1), 26–43 (1999)
10. Vaughn, R.: How Advertising Works: A Planning Model. Journal of Advertising Research 20(5), 27–33 (1986)
11. Wells, J.D., Valacich, J.S., Hess, T.J.: What signal are you sending? how website quality influences perceptions of product quality and purchase intentions. MIS Quarterly 35(2), 373–396 (2011)
12. Wu, S., Lo, C.: The influence of core-brand attitude and consumer perception on purchase intention towards extended product. Asia Pacific Journal of Marketing and Logistics 21(1), 174–194 (2009)

Exploring Human Brands in Online Shopping: An Eye-Tracking Approach

Seong Wook Chae[1], Young Wook Seo[2], and Kun Chang Lee[3,*]

[1] National Information Society Agency, Republic of Korea
[2] Software Engineering Center at NIPA, Republic of Korea
[3] SKK Business School and Department of Interaction Science, Sungkyunkwan University, Seoul 110-745, Republic of Korea
{seongwookchae,seoyy123,kunchanglee}@gmail.com

Abstract. Trust plays a critical role in facilitating transactions in the online shopping environment. Accordingly, various methods have been considered to enhance customer trust. Human branding has received increased attention and played a vital role in business in recent years because it has great impacts on our daily life and consumption. The purpose of this paper is to investigate the effect of applying human brands in an online shopping environment with an emphasis on product type and human brand attachment. The study combines the eye-tracking technique with a self-reported questionnaire to gain a deeper understanding of the effect of human branding in the online shopping process. The results showed that both the product type and level of human brand attachment have significant influences on a customer's visual attention as well as perceived trust towards the product.

Keywords: Human brand, Human brand attachment, Product type, Eye-tracking, Trust.

1 Introduction

Trust is more important in the context of e-commerce than in traditional commerce because online services and products are not typically verifiable immediately. The lack of customer trust towards products and vendors is regarded as an inhibitor in the online shopping environment. Customers report that online stores are impersonal, and there is wide agreement that e-commerce suffers from a lack of customer trust [1-3]. Indeed, research has shown that high levels of customer trust encourage online purchase intentions [4] and help retain online customers [5], while a lack of customer trust is the main reason why individuals do not shop online [6]. Therefore, various methods are applied to build customer trust in the online shopping environment, such as enhancing social presence, providing online reviews, and employing humanoid agents. In the impersonal online environment, using images or an agent that looks like a human is regarded as an effective way to enhance trust.

Human branding has recently received increased attention and played a critical role in business because it has a great influence on our daily life and consumption.

* Corresponding author.

J.-S. Pan, S.-M. Chen, N.T. Nguyen (Eds.): ACIIDS 2012, Part III, LNAI 7198, pp. 10–18, 2012.

Companies are dedicated to managing human brands and building emotional bonds with customers. Feelings linked to attachment are fundamental to strong brand relationships [7], so attachment is an important factor when considering human brands.

Although there are many kinds of products in online shopping malls, they can generally be categorized into two types of products: functional products and symbolic products. While a functional product has utilitarian functions, symbolic products appeal to a customer's affective gratification. The different values of each product lead to different approaches in the decision to make a purchase [8]. Furthermore, these different types of products may be influenced by human brands in an online shopping environment.

The eye-tracking method has recently been used to record and analyze individuals' visual attention by tracing eyesight. It has been employed in various fields such as usability, marketing, cognitive and behavioral psychology, and so on. In this respect, the aim of this study is to employ the eye-tracking technique to examine the effect of applying human brands in the online shopping environment with an emphasis on product type and human brand attachment.

2 Literature Review

2.1 E-Commerce and Efforts to Cope with Its Impediments

Survey researchers argue that the most significant impediments of online shopping are the absence of pleasurable experiences, social interaction, and personal consultation [9]. In addition, the lack of trust in products and vendors is also regarded as an inhibitor in online malls. Accordingly, researchers have attempted to find and explore ways to facilitate e-commerce. One approach is to examine website characteristics that could enhance users' perceptions of social presence, which describes the extent to which a medium is perceived as sociable, warm, personal, or intimate when used to interact with others: socially rich text contents and personalized greetings [2], emotive text and pictures of humans [10], functionalities such as live chat and online reviews [11], and website aesthetics and emotional appeal for the user [12]. Another approach is to investigate the sales assistant functions to improve customers' shopping experiences, such as a software-based product recommendation agent [13], virtual salesperson [3], or avatar sales agent [14].

2.2 Human Brands and Human Brand Attachment

A brand can be broadly defined as a trade marketable visual or verbal piece of information that identifies a product or service. Traditionally, brands have been related to products, services, or organizations, but today researchers appreciate that brands can also be human [15]. The human brand refers "to any well-known persona who is the subject of marketing communication efforts" (Thomson, 2006, p. 104), and has an intangible asset such as a social reputation, image, or credibility. For example, celebrity brands (e.g., Super Junior), athlete brands (e.g., Yuna Kim), and CEO brands (e.g., Steve Jobs) can be thought of as human brands. Today, celebrities are regarded not only as famous entertainers or sports stars in people's minds, but also as a human brand. Accordingly, companies spend great sums each year in an effort to

establish psychological connections between customers and human brands.

Meanwhile, human brand attachment can be referred to as a person's target-specific emotional bond with a human brand [15]. An attachment is a type of strong relationship that people usually first experience as children with their parents. A person immersed in such an emotionally significant relationship normally perceives the relationship partner as irreplaceable [16]. When these types of relationships are experienced in reference to human brands, they are typically referred to as "secondary object" attachments.

2.3 Functional and Symbolic Products

In general, most products have both functional and symbolic dimensions. Some products, however, are basically made for the purpose of instrumental or utilitarian reasons, or for the customers' consummatory affective gratification only [17]. A functional product has the value of pertaining to the utilitarian functions a product can perform (its use). Products differ in the degree to which they are suited to perform their basic utilitarian function [18], such as communication or transportation, but also in terms of quality and features. On the contrary, symbolic products may convey the experiential aspects of consumption, such as customer fantasies, feelings, and fun; the kind of person someone is or wants to be; or to express their (ideal) self-image to themselves and others [18]. The benefits relate to underlying needs for social approval or personal expression and outer-directed self-esteem.

3 Development of Hypotheses

Jillapalli and Wilcox (2010) revealed that strong attachments influence trust and satisfaction [19]. In addition, Thomson (2006) pointed out that strong attachments are predictive of satisfying, trusting, and committed relationships [15]. Hence, we propose the following hypotheses.

Hypothesis 1: Human brand attachment influences customers' visual attention.

Hypothesis 2: Human brand attachment positively influences customers' perceived product trust.

When customers make a decision to choose functional products, they are required to engage in cognitive information processing. By contrast, the purchase motivation of a symbolic product is more influenced by the customer's experiential affect associated with the product than by his or her cognitive information processing. While the choice of functional products depends on a customer's evaluation of product attributes, that of symbolic products depends upon the symbolic factors [8]. Thus, we propose the following hypotheses.

Hypothesis 3: The product type influences customers' perceived product trust.

Hypothesis 4: The product type influences customers' visual attention.

4 Methods

4.1 Pre-test

The human brands and products used in this experiment were chosen by conducting a pretest with surveys and in-depth interviews with 11 participants. First, we listed 23 human brands and 16 products based on the results of interviews with undergraduate and graduate students. Then, we selected the 10 most favorable human brands by investigating the respondents' preferences for each and selected 2 products by conducting a pre-survey with items on product type, with responses to questions on a 5-point scale. A laptop was selected as the functional product ($mean_{functional}$ = 4.82, $mean_{symbolic}$ = 2.24) and perfume as the symbolic product ($mean_{functional}$ = 1.68, $mean_{symbolic}$ = 4.00). In this study, Amazon (www.amazon.com) was chosen as the Internet shopping mall through in-depth interviews.

4.2 Participants

For this experiment, 40 healthy subjects were recruited at a college in South Korea. We received written informed consent from all participants, who were all college students, and each subject was paid 10,000 Korean won for participation. Among the participants, 2 subjects were eliminated from the study because of corrupted data. Data from a total of 38 subjects (16 men and 22 women) were thus employed in this study. About 74% were 21 to 28 years old, and 24% were 20 years or younger. Consequently, the subjects were divided into two groups (high human brand attachment group, n = 16; low human brand attachment group, n = 22).

4.3 Materials and Design

We made a screen with detailed product descriptions consisting of the product, message, and human brand domain for this experiment. The product domain represents the product for advertising, and the message domain is the text presented with the product, including abstractly written sentences about the product concept and a detailed explanation. Finally, the human brand domain uses celebrities as brands with images to enhance consumer interest. Stimuli for the experiment with human brands include a total of 20 types of screen with detailed product descriptions made by combining 10 human brands and 2 product types.

The experimental design used to examine the effect of applying human brands in an online shopping mall environment was a two-factor repeated measure design with two levels for each factor. The first factor of the design was a within-factor of product type, and the second was a between-factor of level of human brand attachment.

4.4 Apparatus

The online shopping mall screen shots were presented on a 19-inch monitor with a resolution of 1024 x 768 pixels. The Tobii Eye Tracker™ (X120) was employed to record participants' eye movement during the experiment, and data were treated by Tobii Studio. Eye Tracker can measure participants' visual attention, which consists of eyeball fixation and saccade. Eyeball fixation shows how long a participant's eyes stay fixed on a certain area and saccade is the momentary movement between eyeball

fixations. Fixations were detected at 100 ms or above, an appropriate cut-off point for tracking eye movements in this experiment [20].

4.5 Procedures

The experimental procedure was as follows. First, before starting this experiment, we asked participants to select their favorite human brand out of the 10 brands listed. Second, a calibration test was conducted to correctly trace each participant's eye movements before starting the experiment. Third, two shopping processes (functional and symbolic) were sequentially displayed to participants as shown in Figure 1 to simulate the real environment for online shopping. For time control of a participant's visual attention measure, each screen was presented for 10 seconds. Lastly, participants were asked to complete the questionnaire and were interviewed.

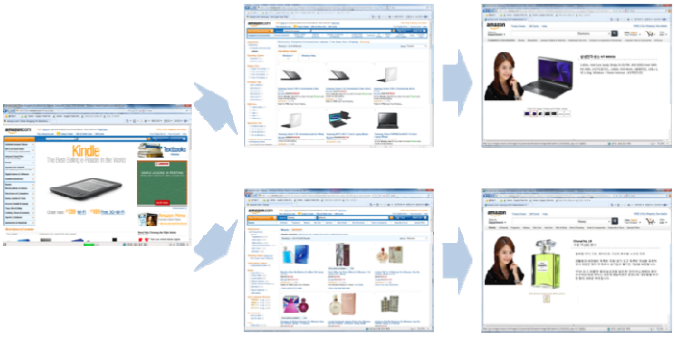


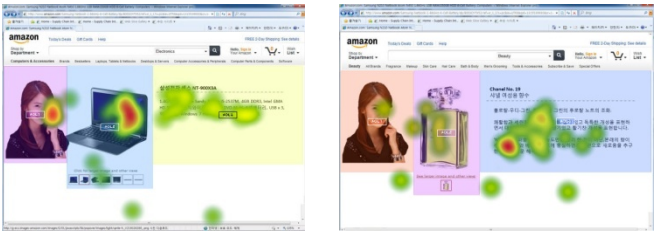
Fig. 1. Experiment process of shopping products (functional / symbolic)

5 Results

5.1 Effects of Product Type and Human Brand Attachment on Visual Attention

Figure 2 is the heat map visualized by Tobii Eye Tracker. It shows the average fixation length of participants, which can be interpreted as the degree of visual attention for each consideration factor, or the area of interest (AOI). With respect to the product AOI (middle part), it is intuitively known that the functional product got more visual attention than the symbolic product.

Repeated measures analysis of variance was employed to statistically verify the heat map results. The eye movement data on fixation length for the product AOI were analyzed to investigate the effect of product type and human brand attachment on visual attention. As shown in Table 1, the main effect of product type was significant ($F(1,36)=3.456$, $p<0.10$), with the functional product (mean=1.651) having significantly longer fixation length than the symbolic product (mean=1.272). In addition, the results showed a marginally significant main effect of human brand attachment ($F(1,36)=3.221$, $p<0.10$). The group of high human brand attachment (mean=1.705) had relatively longer fixation length than the low human brand attachment group (mean=1.217), thus validating H1 and H2, respectively.



(a) Functional product (b) Symbolic product

Fig. 2. Heat map measured by eye-tracking

Table 1. ANOVA on product area

Source	Sum of Squares	df	Mean Square	F
Product type (A)	2.664	1	2.664	3.456^{+}
Human brand attachment (B)	4.416	1	4.416	3.221^{+}
A x B	.485	1	.485	.629

**NOTE: Statistically significant at $^{+}p<0.10$*

5.2 Effects of Product Type and Human Brand Attachment on Perceived Product Trust

The results showed two main effects and no interaction effect as shown in Table 2 and Figure 3, which indicated similar results with the case of visual attention. The main effect of product type was significant ($F(1,36) = 8.922$, $p<0.01$), with functional product (mean = 4.968) having significantly greater perceived trust towards the product as compared with the symbolic product (mean = 4.379). Moreover, results revealed a marginally significant main effect of human brand attachment ($F(1,36)=3.994$, $p<0.10$). The group of high human brand attachment (mean=4.968) had relatively higher perceived product trust as compared with the low human brand attachment group (mean = 4.379). These results support H3 and H4, respectively.

Table 2. ANOVA of perceived product trust

Source	Sum of Squares	df	Mean Square	F
Product type (A)	12.668	1	12.668	8.922**
Human brand attachment (B)	6.445	1	6.445	3.994^{+}
A x B	.087	1	.087	.062

**NOTE: Statistically significant at $^{+}p<0.10$, $^{**}p<0.01$*

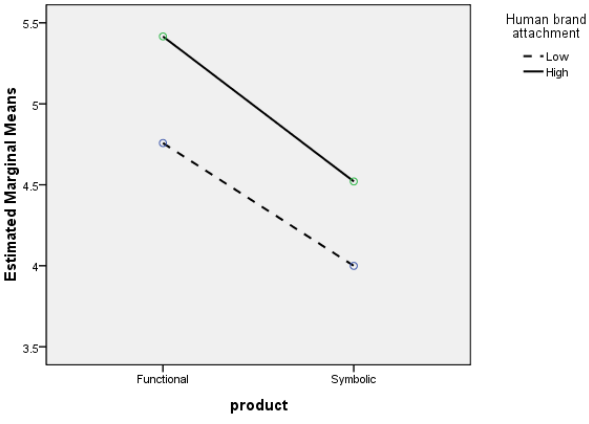


Fig. 3. Perceived trust towards product

6 Discussion

In this study we investigated an effective way to use human brands in the online shopping context with eye tracking technology. The results of the current study led to the following findings. First, there is a significant difference in visual attention to functional products and symbolic products. Specifically, the fixation length when viewing a functional product is much longer than that when viewing a symbolic product. In other words, customers may look at every attribute of the functional product in detail when processing information during the purchase, but intuitively perceive the image and meaning of a symbolic product in a more holistic way. Another difference is found in the level of human brand attachment. Customers with relatively high human brand attachment have a long fixation length for products compared with customers who have low attachment. This could be because customers with strong human brand attachment for the product presented on the screen are more likely to have greater visual attention to the product when there are detailed product descriptions. Second, the results reveal that the product type influences a customer's perceived trust regarding a product. According to our results, customers have more trust for functional products than for symbolic products. In general, individuals are more likely to build trust in tangible, specific things as compared with intangible, abstract, and non-specific things. Furthermore, symbolic products are generally used to express one's self-image [18, 21-23], so purchasing the wrong symbolic products could result in high social and psychological risk. As compared with symbolic products, functional products have a relatively lower level of social and psychological risk. Therefore, considering the potential risk of purchasing the wrong product, it appears that building trust towards a symbolic product is harder than building trust towards a functional product. Additionally, the results indicate that there is a

significant difference in perceived product trust between the levels of human brand attachment. Specifically, when customers have higher attachment towards a human brand, they build more trust towards the product. That is, strong attachments influence trust [19] and are predictive of trusting, committed relationships [15], as we expected. Finally, in terms of product type and the level of human brand attachment, the results of perceived product trust are very much like those of a customer's visual attention. In other words, when customers fix their eyes on a certain product for a longer time, they also perceive more trust towards the product. It is highly probable that looking at something favorably for a certain amount of time might bring trust up in one's mind.

One of this paper's contributions is employing a multi-method approach, a self-reported questionnaire and eye-movement data, to gain a deeper understanding of the data, especially when observing complex phenomenon, for which customers themselves may not be aware of their reactions. Nonetheless, this research has limitations of a relatively small sample size and the use of a sample of college students, which may not be representative of the general population. Future researchers could utilize experiments to include and analyze two other areas of the screen for detailed product descriptions, such as the human brand area and message area, which could result in additional insights and generalized experimental results.

7 Concluding Remarks

Considering human brands as a facilitation method for online shopping, it would be favorable to use human brands for which customers are more likely to have attachment in order to attract a customer's visual attention and build customer trust towards the product. In addition, it would be helpful to consider product type, such as functional product and symbolic product, in order to effectively enhance product trust.

Acknowledgment. This research was supported by the World Class University (WCU) program through the National Research Foundation of Korea, funded by the Ministry of Education, Science and Technology, Republic of Korea (Grant No. R31-2008-000-10062-0).

References

1. Brynjolfsson, E., Smith, M.: Frictionless Commerce? A Comparison of Internet and Conventional Retailers. Management Science 46, 563–585 (2000)
2. Gefen, D., Straub, D.W.: Managing user trust in B2C e-Services. e-Service Journal 2, 7–24 (2003)
3. Komiak, S.Y.X., Weiquan, W., Benbasat, I.: Trust Building in Virtual Salespersons Versus in Human Salespersons: Similarities and Differences. e-Service Journal 3, 49–63 (2004)
4. Gefen, D.: e-Commerce: The role of familiarity and trust. Omega 28, 725–737 (2000)

5. Reichheld, F., Schefter, P.: E-loyalty: your secret weapon on the web. Harvard Business Review 78, 105–113 (2000)
6. Hoffman, D.L., Novak, T.P., Peralta, M.A.: Building consumer trust online. Communications of the ACM 42, 80–85 (1999)
7. Fournier, S.: Consumers and Their Brands: Developing Relationship Theory in Consumer Research. Journal of Consumer Research 24, 343–373 (1998)
8. Hirschman, E.C.: Symbolism and Technology as Sources for the Generation of Innovations. Advances in Consumer Research 9, 537–541 (1982)
9. Barlow, A.K.J., Siddiqui, N.Q., Mannion, M.: Development in Information and Communication Technologies for Retail Marketing Channels. International Journal of Retail and Distribution Management 32, 157–163 (2004)
10. Hassanein, K., Head, M.: The impact of infusing social presence in the web interface: an investigation across different products. International Journal of Electronic Commerce 10, 31–55 (2006)
11. Cyr, D., Hassanein, K., Head, M., Ivanov, A.: The role of social presence in establishing loyalty in e-Service environments. Interacting with Computers 19, 43–56 (2007)
12. Cyr, D., Head, M., Larios, H., Pan, B.: Exploring Human Images In Website Design: A Multi-Method Approach. MIS Quarterly 33, 539–566 (2009)
13. Qiu, L., Benbasat, I.: Evaluating Anthropomorphic Product Recommendation Agents: A Social Relationship Perspective to Designing Information Systems. Journal of Management Information Systems 25, 145–181 (2009)
14. Holzwarth, M., Janiszewski, C., Neumann, M.M.: The Influence of Avatars on Online Consumer Shopping Behavior. American Marketing Association 70, 19–36 (2006)
15. Thomson, M.: Human brands: Investigating antecedents to consumers' strong attachments to celebrities. Journal of Marketing 70, 104–119 (2006)
16. Hazan, C., Shaver, P.R.: Attachment as an Organizational Framework for Research on Close Relationships. Psychological Inquiry 5, 1–22 (1994)
17. Batra, R., Ahtola, O.T.: Measuring the Hedonic and Utilitarian Sources of Consumer Attitudes. Marketing Letters 2, 159–170 (1990)
18. Strahilevitz, M., Myers, J.G.: Donations to Charity as Purchase Incentives: How Well They Work Depend on What You Are Trying to Sell. Journal of Consumer Research 24, 434–446 (1998)
19. Jillapalli, R.K., Wilcox, J.B.: Professor Brand Advocacy: Do Brand Relationships Matter? Journal of Marketing Education 32, 328–340 (2010)
20. Inhoff, A.W., Radach, R.: Definition and computation of oculomotor measures in the study of cognitive processes. In: Underwood, G. (ed.) Eye Guidance in Reading, Driving and Scene Perception, pp. 29–53. Elsevier, New York (1998)
21. Belk, R.W.: Possessions and the Extended Self. Journal of Consumer Research 15, 139 (1988)
22. Holbrook, M.B., Hirschman, E.C.: The experiential aspects of consumption: consumer fantasies, feelings, and fun. Journal of Consumer Research 9, 132–140 (1982)
23. Solomon, M.R.: The Role of Products as Social Stimuli: A Symbolic Interactionism Perspective. Journal of Consumer Research 10, 319–329 (1983)

Task Performance under Stressed and Non-stressed Conditions: Emphasis on Physiological Approaches

Nam Yong Jo[1], Kun Chang Lee[2,*], and Dae Sung Lee[3]

[1,3] SKK Business School, Sungkyunkwan University,
Seoul 110-745, Republic of Korea
[2] SKK Business School
WCU Professor at Department of Interaction Science
Sungkyunkwan University, Seoul 110-745, Republic of Korea
{namyong.jo,kunchanglee,leeds1122}@gmail.com

Abstract. By using a physiological approach, we examined performance in the Minesweeper[R] game. In this experiment, we measured subjects' Galvanic Skin Response (GSR) and electrocardiogram (ECG) during game play. We divided subjects into two groups, one of which was exposed to two types of manipulated stress. Additionally, a questionnaire was given to the subjects in order to measure perceived stress. We investigated how much stress each group endured by measuring physiological data and by administering the perceived stress scale (PSS). The results showed that there was no difference for multi-relational performance between the control group and the experimental group. For future studies of multi-relational performance under stress, we suggest that researchers should consider other factors that might influence stress and multi-relational performance.

Keywords: stress, game, performance, physiological signals, GSR, ECG.

1 Introduction

Psychological, organizational, and educational literatures have all examined the relationship between stress and performance. This study used a physiological approach to analyze the effect of stress on performance. For this experiment, we subjected the experimental group to two stress manipulations, threat of time pressure and performance feedback. The physiological measures used in our experiment were the Galvanic Skin Response (GSR) and electrocardiogram (ECG). A total of 32 subjects participated in this study. Their mean age was 22 years. Half of the subjects were randomly assigned to the stressed group, the other half were controls (stressed group, n = 16; non-stressed group, n = 16). The subjects were instructed to start meditating for seven minutes in order to acquire baseline data from GSR and ECG electrodes. After that, the subjects were asked to play the Minesweeper[R] game for seven minutes. During this period, GSR and ECG data were recorded. After the

* Corresponding author.

J.-S. Pan, S.-M. Chen, N.T. Nguyen (Eds.): ACIIDS 2012, Part III, LNAI 7198, pp. 19–26, 2012.

physiological experiment, the subjects were requested to complete a questionnaire, which consisted of stress-related questions.

The data analysis explored whether or not our stress manipulations coincided with a) self-reported stress, as measured by the Perceived Stress Scale [PSS] and b) the physiological measures of GSR and ECG. We also investigated under which stress condition game performance was better.

2 Theoretical Background

2.1 Stress and Performance

There has been considerable controversy about the influence of stress on performance. Some researchers insist that stress has a negative influence on performance; others maintain that the effect of stress on performance is positive [1]. According to prior studies, this inconsistency seems to be rooted in two major findings. First, the arousal associated with stress appears to increase performance up to a certain point, but thereafter, as arousal continues to increase, performance declines. Second, there are forms of stress that are referred to as "challenging stress" (e.g., demands of work); these are positively related to performance. There is also "hindrance stress" (e.g., role ambiguity; role conflict), which is assumed negatively related to performance.

For the purpose of our study, we define stress as challenging stress, or stress that has reached a level at which it exhibits positive effects on cognitive performance. Lepine et al. [14] maintain that, although hindrance stress is negatively related to performance, challenging stress promotes motivation to learn and influences performance positively. Thus, in our experiment, we used two challenging stress manipulations (time pressure and performance feedback) with a specific group as they played the Minesweeper^{R} game.

2.2 Minesweeper^{R} and Learning Task

In previous studies, the Minesweeper^{R} game has been used to illustrate the complexity of a multi-relational learning task [2]. Minesweeper^{R} has two major aspects that can be used to describe a user's learning task. First, one realizes the complexity of the game by calculating an estimate for the size of its search space. Consider a 9 what? × 9 board with M = 10 mines (see Fig. 1). At the beginning of the game, the player has 81 tiles from which to uncover tile. In Minesweeper^{R}, there are situations that can be "solved" with nontrivial reasoning. For example, consider Fig. 1 (left) where the only available information about the board status is the numbers. After careful analysis one finds the squares with a mine (see Fig. 1, right) and those that do not contain a mine, one realizes that a square with a flag is a mine, and the state of the blank tiles cannot be determined if one does not know how many mines are hidden in the board. There are other Minesweeper^{R} situations where the information available is not sufficient to identify a safe square or one with a mine, as in Fig. 2, and the best option available to the player is to make an informed guess, i.e., a guess that minimizes the risk of being

blown up by uncovering a mine. In this study, we consider the learning task in Minesweeper[R] to be the deduction of rules to identify all the *safe squares*[1] and squares with a mine that can be deduced given a board's state[2].

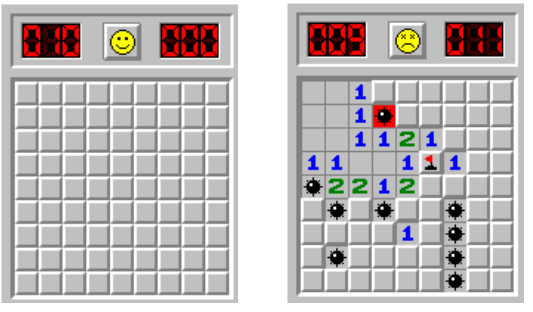


Fig. 1. Left: the initial status of the Minesweeper, Right: the lost status of the Minesweeper (flag and bomb symbolize the place of bomb)

2.3 Assessment of Stress

Psychological Assessment

A stress response can be measured and evaluated in terms of perceptual, behavioral, and physical responses. The evaluation of perceptual responses to a stressor involves subjective estimations and perceptions. Self-report questionnaires are the most common instruments used to measure stress [8]. Representative measures are the Perceived Stress Scale (PSS) [7], the Life Events and Coping Inventory (LECI) [10], and the Stress Response Inventory (SRI) [11]. In this study we used the PSS.

Physiological Assessment

The physical response to stress has two components: a physiological response indicative of central-autonomic activity and a biochemical response involving changes in the endocrine and immune systems [8]. Stress induces a change in autonomic functioning [15]. It affects blood pressure and heart rate, reflecting a predominance of sympathetic nervous system (SNS) activity [13]. Heart rate variability (HRV) is the beat-to-beat variation in heart rate, and it has recently been used as a biomarker of Autonomic Nervous System (ANS) activity associated with mental stress [16]. HRV analysis is generally divided into two categories: time-domain and frequency-domain methods. Time-domain analysis of HRV involves quantifying the mean or standard deviation of RR intervals. Frequency-domain analysis involves calculating the power of the respiratory-dependent high frequency (HF) and low frequency (LF) components of HRV. In this study, we selected the standard deviation of RR intervals (SDNN) and LF/HF ratio as ECG information. Mental stress is reported to evoke a decrease in the high-frequency component and an increase in the low-frequency component of HRV [3]. Therefore, LF/HF ratio increases if mental stress occurs. On the other hand, a decrease of SDNN is also related to metal stress.

GSR is a measure of the electrical resistance of the skin. A transient increase in skin conductance is proportional to sweat secretion [9]. When an individual is under mental stress, sweat-glands are activated; this increases skin conductance. Because the sweat glands are also controlled by the SNS, skin conductance acts as an indicator for sympathetic activation due to the stress reaction.

3 Method

3.1 Participants

Thirty-seven healthy subjects recruited from the undergraduate student population at a Korean university participated in this experiment. Prior to the experiment, the subjects were given written and verbal information explaining the experimental procedures. We confirmed through interviews that none of the subjects used medication for hypertension or any other cardiovascular disease and they were all free of any nervous or other psychological disorder. We received written informed consent from all participants and each subject was paid 20,000 Korean won for his or her participation. Among them, six subjects with corrupted data were eliminated from the study. A total of 32 subjects (23 men and 9 women) were employed in this study. The mean age was 22 years (range of 18-26 years). The subjects were randomly divided into two groups (stressed group, n = 16; non-stressed group, n = 16).

3.2 Experimental Procedure

Before the experiment, the subjects were asked if they knew how to play the Minesweeper^{R} game. In those cases in which a subject did not know the game, they were instructed to practice it for 15 minutes to become accustomed to how to play. Then they were instructed to cleanse their hands and remove all accessories from their bodies. Next, the subjects were asked to sit comfortably and keep their left hand still when the experiment started. Each subject was asked to attach two GSR electrodes to the index and middle fingers of their left hand and to place three ECG electrodes on their chest and abdomen. In this experiment we used a Biopac MP100 series for recording and an AcqKnowledge 4.1 for analysis of physiological data. After GSR and ECG signals showed normal waves, the subjects were instructed to start meditating for seven minutes in order to acquire baseline data from the GSR and ECG electrodes. Finally, they were asked to play the Minesweeper^{R} game for 7 minutes. In the course of task implementation, GSR and ECG signals were measured for both the stressed and non-stressed groups. In addition, stress manipulations were inserted into the stressed group. Their performances were recorded in the form of the number of winning and losing games and the time spent playing. After the physiological experiment, the subjects were requested to complete the questionnaire on perceived stress. Fig.2. shows how to record the subjects' scores and the physiological data.

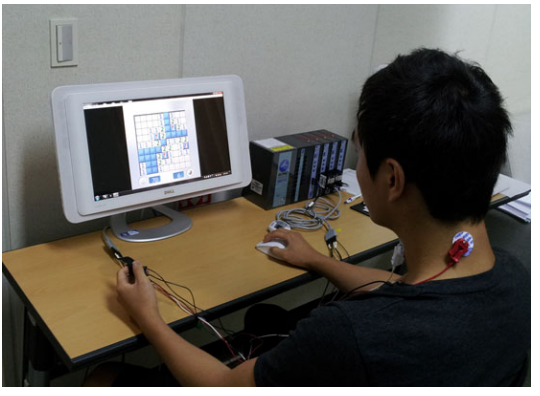


Fig. 2. Results of the structural model

Performance Task
[2] also used the Minesweeper[R] game to assess performance. In their study, subjects' performance was measured only as accuracy, not speed. [5] used the number of correctly placed flags per second (Pmines) to measure performance. We adopted both of these measures, thus normalizing both the number of games won and the average time to win a game. We then integrated both variables into a performance index using Principal Component Analysis (PCA).

Stress Manipulation
This experiment used two stress manipulations in the stressed group (threat of shock and performance feedback). These two stress manipulations had been used by [4]. In the case of the threat-of-shock manipulation (no shock was ever actually delivered), the subjects were informed that they might receive an "unpleasant but not painful" electrical shock through the electrodes attached to their bodies. They were instructed that the possibility that they would receive a shock was dependent upon their performance in comparison to past subjects. In the performance-feedback manipulation, the subjects were told that they must be lacking in creativity and that they were less creative than previous participants. These manipulations were implemented according to a fixed pattern, independent of actual performance.

Questionnaire Survey
In order to compare manipulated stress in the experiment with perceived stress, we selected PSS as our survey tool [6, 7]. The PSS measures the degree to which situations are considered stressful, by addressing events experienced beforehand. It was designed to quantify how unpredictable and uncontrollable adults find their lives. Each item in our survey was measured on a seven-point Likert scale, with answers ranging from ''strongly disagree" to ''strongly agree." The items in the survey were developed by adopting existing measures validated by other researchers.

3.3 Statistics

For physiological data and performance assessment, the differences between the stressed and non-stressed groups were analyzed with the Mann-Whitney U Test. This

test was performed to test the null hypothesis that the stressed group was not different from the non-stressed group. The results from the Mann-Whitney U Test are presented with the p-value. Statistical significance was assumed for $P < 0.05$. We investigated the two types of performance ratings for each group with the Wilcoxon signed ranks test. Finally, we examined the relationship between stress and performance through descriptive statistics.

4 Results

4.1 Differences between Stressed Group and Non-stressed Group

Physiological Data and Performance Assessment

The relationship between manipulated stress and physiological data (Normalized Δ GSR, Δ SDNN, and Δ LF/HF ratio) was investigated using the Mann-Whitney U Test. This test is one of the most powerful nonparametric tests, and is a most useful alternative to parametric tests when the researcher wishes to avoid the t test's assumptions or when sample sizes are relatively small [14]. Although there was no significant difference between the stressed and non-stressed groups for normalized Δ GSR and Δ SDNN, as shown in Table 1, we confirmed that the stressed group had a significantly higher Δ LF/HF ratio than the non-stressed group. On the other hand, performance as assessed by game results was not statistically different between the two groups.

Table 1. Mann-Whitney U Test Results for Physiological Signals and Learning Performance

Group	N	*Normalized* Δ GSR		*Normalized* Δ SDNN		*Normalized* Δ LF/HF ratio		Performance	
		Mean	SD	Mean	SD	Mean	SD	Mean	SD
Stress	16	-0.029	0.215	0.161	0.418	0.993	1.292	0.308	1.741
Non-stress	16	-0.045	0.234	0.471	0.789	0.109	0.642	-0.352	1.798
Total	32	-0.037	0.221	0.316	0.641	0.551	1.099	-0.022	1.773
Two-Tailed Probability		0.624		0.122		**0.042** *		0.344	

* Statistically significant at $p < 0.05$

Self-reported Stress

Statistically significant differences between the two groups were observed for perceived stress and self-reported stress. This result shows that our manipulation of stress was well controlled in the experiment and that it discriminated the stressed group from the control group. While we did not verify the difference between the two groups for performance through the Mann-Whitney U Test, we made sure that the stressed group had more perceived stress than the control, as shown in Table 2. In this view, the subjects are thought to be properly divided. Therefore, we would confirm which group could separately explain performance by comparing physiological and self-reported stress.

Table 2. Mann-Whitney U Test Results for Physiological Signals and Self-reported Stress

Group	N	*Normalized* Δ GSR		*Normalized* Δ SDNN		*Normalized* Δ LF/HF ratio		Self-reported Stress	
		Mean	SD	Mean	SD	Mean	SD	Mean	SD
Stress	16	-0.029	0.215	0.161	0.418	0.993	1.292	4.475	1.012
Non-stress	16	-0.045	0.234	0.471	0.789	0.109	0.642	3.213	1.018
Total	32	-0.037	0.221	0.316	0.641	0.551	1.099	3.844	1.187
Two-Tailed Probability		0.624		0.122		**0.042***		**0.007** **	

* Statistically significant at $p < 0.05$

** Statistically significant at $p < 0.01$

5 Discussion

There has not been an abundance of research regarding the effects of stress on performance. Thus, we assessed those effects to see if we could demonstrate different relationships with performance between stressful and non-stressful conditions. Although our hypotheses regarding stress's effect on performance were not supported by our statistical results, we suggest it should not be concluded that there are no such effects, as there were some limitations in our experiment's design. In our opinion, these limitations deserve to receive attention in future research, because an agreement has not yet been reached on the relationship between stress conditions and performance. We suggest further work on two main design issues in the stress-performance experiments. First, the stressed condition should be more concrete and narrow. In our experiment, we gave subjects challenging stress in the form of time pressure. Compared to other subjects, it was not clear whether ours actually felt challenged, which might also differ depending on the subjects' personalities and perceptions of challenging stress.

Moreover, it has not been verified that time pressure and apparent comparison to other players constitute homogeneous stress. Third, as in previous research, mediating constructs should be considered. For example, although relationships between the stress variables and motivation to learn can be examined, existing theories describing this have not been much studied, nor has an agreement about these relationships been reached [17]. Finally, to reduce unnecessary variability, subjects should be selected based on an existing familiarity with the game. When subjects are not fully familiar with the learning tool (i.e., Minesweeper^{R} in this study), simply learning the game produces stress. Consequently, the performance deviation is associated with skill, rather than the experimental condition. That is, although challenging stress may be motivational and promote performance, the relationships could not be demonstrated statistically in this study. We recommend that the factors we discussed be controlled strictly in future research.

Acknowledgments. This study was supported by WCU (World Class University) program through the National Research Foundation of Korea funded by the Ministry of Education, Science and Technology (Grant No. R31-2008-000-10062-0).

References

1. Doris, F., Sabine, S.: Rethinking the Effect of Stress: A longitudinal Study on Personal Initiative. Journal of Occupational Health Psychology 7, 221–234 (2002)
2. Castillo, L., Wrobel, S.: Learning Minesweeper with Multi-relational Learning. In: Proceedings of the International Joint Conference on Artificial Intelligence, pp. 533–538 (2003)
3. Bernardi, L., Wdowczyk-Szulc, J., Valenti, C., Castoldi, S., Passino, C., Spadacini, G., Sleight, P.: Effects of Controlled Breathing, Mental Activity, and Mental Stress with or without Verbalization on Heart Rate Variability. Journal of the American College of Cardiology 35, 1462–1469 (2000)
4. Bogdan, R., Pizzagalli, D.A.: Acute Stress Reduces Reward Responsiveness: Implications for Depression. Biological Psychiatry 60, 1147–1154 (2006)
5. Ivanov, S.V.: Theoretical and Experimental Models in Physiological and Psychological Research of Pattern Perception and Recognition in Humans. Pattern Recognition and Omage Analysis 19, 114–122 (2009)
6. Cohen, S., Williamson, G.: Perceived Stress in a Probability Sample of the United States. In: Spacapan, S., Oskamp, S. (eds.) The Social Psychology of Health. Sage, Newbury Park (1988)
7. Cohen, S., Kamarck, T., Mermelstein, R.: A global Measure of Perceived Stress. Journal of Health and Social Behavior 24, 386–396 (1983)
8. Cohen, S., Kessler, R., Gordon, L.: Measuring Stress - A Guide for Health and Social Scientists. Oxford University Press (1997)
9. Darrow, C.: The Rationale for Treating the Change in Galvanic Skin Response as a Change in Conductance. Psychophysiology 1, 31–38 (1964)
10. Dise-Lewis, J.E.: The Life Events and Coping Inventory: An Assessment of Stress in Children. Psychosomatic Medicine 50, 484–499 (1988)
11. Koh, K., Park, J., Kim, C., Cho, S.: Development of the Stress Response Inventory and its Application in Clinical Practive. Psychosomatic Medicine 63, 668–678 (2001)
12. LePine, J.A., Podsakoff, N.P., LePine, M.A.: A Meta-analytic Test of the Challenge Stressor-hindrance Stressor Framework: An Explanation for Inconsistent Relationships among Stressors and Performance. Academy of Management Journal 48, 764–775 (2005)
13. Ritvanen, T., Louhevaara, V., Helin, P., Vaisanen, S., Hanninen, O.: Responses of the Autonomic Nervous System during Periods of Perceived High and Low Work Stress in Younger and Older Female Teachers. Applied Ergonomics 37, 311–318 (2005)
14. Siegel, S., Castellan, N.J.: Nonparametric Statistics for the Behavioural Sciences. McGraw-Hill, New York (1988)
15. Van der Kar, L.D., Blair, M.L.: Forebrain Pathways Mediating Stress Induced Hormone Secretion. Frontiers in Neuroendocrinology 20, 41–48 (1999)
16. Zhong, X., Hilton, H.J., Gates, G.J., Jelic, S., Stern, Y., Bartels, M.N., DeMeersman, R.E., Basner, R.C.: Increased Sympathetic and Decreased Parasympathetic Cardiovascular Modulation in Normal Humans with Acute Sleep Deprivation. Journal of Applied Physiology 98, 2024–2032 (2005)
17. Perrewe, P.L., Zellars, K.L.: An examination of attributions and emotions in the transactional approach to the organizational stress process. Journal of Organizational Behavior 20, 739–752 (1999)

Characteristics of Decision-Making for Different Levels of Product Involvement Depending on the Degree of Trust Transfer: A Comparison of Cognitive Decision-Making Criteria and Physiological Responses

Min Hee Hahn[1], Do Young Choi[2], and Kun Chang Lee[3,*]

[1] SKK Business School, Sungkyunkwan University,
Seoul 110-745, Republic of Korea
[2] Solution Business Unit, LG CNS Co., Ltd,
Seoul 100-725, Republic of Korea
[3] SKK Business School, WCU Professor at Department of Interaction Science
Sungkyunkwan University, Seoul 110-745, Republic of Korea
{minheehahn,dychoi96,kunchanglee}@gmail.com

Abstract. This research verified the match rate between cognitive decision-making criteria and the physiological reaction on a website using Eye-Tracker, which measures a user's visual attention. Specifically, we explored the match rate between cognitive decision-making and the physiological reaction depending on the degree of product involvement and the degree of trust transfer through an experiment and survey. The verification results show that, for the involvement of products (high/low involvement), the match rate between the fixation length and the cognitive criteria used in decision-making for the high involvement product was higher than that with the low involvement product, and the difference in the match rate was statistically significant. However, in the aspect of trust transfer, the difference in the match rate between the high and low trust transfer groups was not statistically significant.

Keywords: Cognitive decision-making, Product involvement, Trust transfer, Eye-tracking.

1 Introduction

An individual with limited information processing ability makes decisions in a quasi-rational way, which compounds simplifying heuristics and rational analysis [4]. That is, in most cases, individuals make judgments and decisions in a semi-reasonable way that combines the intuition from their accumulated experience and their own decision-making rules. However, there are always flaws no matter what kind of decision-making mechanism we use. In addition, people have a tendency to rely on a few specific pieces of information or clues when processing information [5]. In the case

* Corresponding author.

J.-S. Pan, S.-M. Chen, N.T. Nguyen (Eds.): ACIIDS 2012, Part III, LNAI 7198, pp. 27–36, 2012.

where much information is provided, not all of that information is used in the decision-making process. Rather, a user's efficient decision rules are applied only to selected information [13].

This research explored the relationship between cognitive decision-making criteria and actual physiological reactions in the decision-making process. More specifically, the research question asks if there is a relationship between the cognitive decision-making criteria and the actual physiological reaction (e.g., eye movement), and we analyzed the difference between groups with varying degrees of trust transfer and product involvement. For this research, we conducted an experiment that measures the fixation length using Eye-Tracker, and also collect data through a survey.

In this paper, we first consider previous studies related to product involvement and trust transfer. Next, we explain the research methods for the experiment and the survey, and provide the results of the experiment and data analysis. Finally, we suggest implications of this research based on the results and provide directions for future research.

2 Previous Studies

2.1 Trust Transfer

Online trust is an important factor in e-business activities such as B2B and B2C. Corritorea *et al.* [3] pointed out that trust is the core success factor in the online environment. In this research, we define online trust as trust for the purchasing individual or business in the e-business, with activities carried out through an electronic mechanism such as a website or electronic network.

This type of research on trust is expanding to research on "trust transfer" following the expansion of linked business activities between online and offline businesses. Trust transfer means that a particular entity relies on another particular entity, although where the transferred trust comes from is not known [11]. Lee *et al.* [8] suggested the recognition of trust transfer model, arguing that trust transfer occurs when individuals change their transaction channel from offline store A to online website B. For example, customers who use an offline bookstore have higher trust in the same company's online bookstore as compared with a pure online bookstore (that does not have an offline store). In conclusion, the trust in A is transferred to trust in B.

In this research, we analyze whether trust in the website transfers to trust in information provided by the website and explore the relationship between the cognitive decision-making process and the physiological reaction depending on the degree of trust transfer.

2.2 Product Involvement

One important variable that affects the amount of information searched for by a customer is product involvement. The general conclusion in existing research is that the process of searching for information in purchase decisions differs by the level of involvement. We expect that the information search process will differ depending on the level of involvement and whether the online or offline information also affects the choice of the product purchase channel.

Rothschild and Gaidis [10] argued that, under high involvement, as individuals improve their behavioral ability by interpreting information today based on past information, the information search activities increase. In addition, the evaluation becomes more complicated as the individual tries to use all of the information gathered in the decision-making process. Robertson *et al*. [9] reported that for consumer behaviors such as purchase-related information searches and processing, the ease with which the attitude changes also depends on the level of involvement.

On the other hand, the general conclusion of research on the characteristics and involvement of the online environment is that the more highly involved customers are, the more positive their attitude will be toward online shopping. Cho *et al.* [2] analyzed the ability of online customers to judge the quality of products. The results of the empirical analysis indicate that the effects of the perceived price, search information efforts, and involvement differ by the product category. They argued that the information search effort of customers is related to the level of involvement in the cyber space. In this research, we investigate information search efforts with a focus on high and low involvement products by observing an individual's eyesight movement and compare this with the cognitive decision-making criteria.

3 Research Questions and Hypotheses

Considering previous studies and other materials, in the present study we investigate if there is a difference in the match between the reaction from the cognitive decision-making criteria and the physiological reaction in online shopping. According to the capacity theory of attention, cognitive resources affect humans' cognitive behavior, and the cognitive resources of people are limited. These limited resources can be allocated to a number of activities and the allocation for each activity depends on factors such as the assigned task and given stimuli [7].

In this context, if the involvement level of a product is high, the customer considers many factors in the decision-making process, which leads to an increase in content about differences between products. As a result, customers require more information to make a decision about the differences between products. In contrast, low involvement customers put little effort in the search for information that can explain the differences between products and handle information through the surrounding channels [1]. Previous researchers have presented empirical research about the dependence of the cognitive decision-making process on the degree of product involvement, but there is limited research on this topic in relation to actual physiological data. Thus, we propose the following hypotheses.

H1: The match rate between the cognitive decision-making criteria and the physiological reaction differ by the degree of product involvement.

H1-1: In the high degree of trust transfer group, the match rate between the cognitive decision-making criteria and the physiological reaction differ by the degree of product involvement when products are bought in the online shopping environment.

H1-2: In the low degree of trust transfer group, the cognitive decision-making criteria and the physiological reaction differ by the degree of product involvement when products are bought in the online shopping environment.

In online transactions, where people do not see each other face-to-face, trust becomes more important. In addition, human beings have the intrinsic tendency to achieve cognitive equilibrium and therefore tend to have psychological easement by maintaining consistency between their attitudes and beliefs [6]. Therefore, we can deduce that trust in an online website can be transferred to trust in the information provided on that website. We can further deduce that, depending on the degree of trust transfer, there may be some relationship between the decision-making criteria and the physiological reaction. Therefore, we propose the following hypotheses.

H2: The match rate between the cognitive decision criteria and the physiological reaction will differ by the trust transfer degree when products are purchased online.

H2-1: In the group looking at high involvement products online, the match rate between the cognitive decision-making criteria and the physiological reaction will differ by the trust transfer degree.

H2-2: In the group looking at low involvement products online, the match rate between cognitive decision-making and the physiological reaction will differ by the trust transfer degree.

4 Research Methods

4.1 Experimental Design and Procedures

We collected data in September 2011 through a survey of 70 students attending universities in Seoul. Each participant was allocated to Experimental Group 1 or 2 and joined the experiment about two kinds of stimulus (Table 1).

Table 1. Experimental Group and Stimulus

Group	Stimulus classification	Experimental Stimulus	Description
1	A	The purchase of the high involvement product in the well-known website	Purchasing of the second-hand car in the SK Encar site.
	B	The purchase of the low involvement product in the relatively unknown website	Purchasing of the second hand book in the Second hand book love website
2	C	The purchase of the high involvement product in the relatively unknown website	Purchasing of the information for the second hand car in the Automax website
	D	The purchase of the low involvement product in the well-known website	Purchasing of the information of the second hand book in the Kyobo bookstore website

The length of the experiment was about 20 minutes. After the experiment, participants were asked to complete the questionnaire about the reliability of the website and information provided for the products they looked at. First, by using the Eye-Tracker program developed by Tobbi in Sweden, we collected data on subjects'

eye movement. This apparatus allows the experiment to be conducted easily and conveniently because it has a sight tracking device and requires no other materials.

Eye-Tracker was used to measure the visual attention of users. In this research, we defined the fixation length, which indicates how long the participant's eye-sight was focused on the product information location in the online shopping mall, as the visual attention of participants. In addition, to explore the cognitive decision-making criteria after the eye-tracking was recorded, we asked participants to write down three important factors considered in the purchase decision-making process through another survey. This was used to gather information on the cognitive decision-making criteria used by individuals for the high and low involvement products. To measure the physiological reaction to the stimuli, we measured the fixation length for each Area of Interest (AOI) for the third stimulus, as described in Fig. 1.

Fig. 1. An Example of Eye-tracking Result

5 Results

5.1 Data Analysis

The purpose of this research is to classify people based on their degree of trust transfer (high and low groups) and compare the match rate between the cognitive decision-making criteria and the physiological reaction depending on the degree of product involvement. Thus, we group the participants based on the degree of trust transfer. In general, the risk that the customers perceive about online shopping is much greater than that perceived for offline shopping [12], so participants are very likely to carry out transactions on a website with high reliability to minimize their risk. If customers have an amicable attitude toward an online website, they will also positively accept the information about the products provided by that website. The customers' trust that forms from the website affects the information provided, which can be explained by the "trust transfer". Trust transfer is derived from the concept that customers use clues from the past in purchase decisions to minimize the transaction risks of new channels in the purchasing environment, which is becoming more diverse and complicated [11].

To judge the degree of trust transfer, we collected survey data from the subjects, asking them to state their opinion about 'trust in the website itself (A)' and 'trust in the product information (B)' on a 7-point Likert scale (1 strongly disagree, 7 strongly agree). We regarded those having a score higher than 4 points as the group with a high degree of trust and those with a score less than 4 points as the group with a low degree of trust. If the result from (A) and (B) was high/high or low/low, we regarded that the trust in the website itself was transferred to trust in the product information and judged the degree of trust transfer to be high. If the result from (A) and (B) was high/low or low/high, we concluded that the degree of trust transfer was low, because the trust in the website itself was different from trust in the product information. Table 2 presents the results of the survey questions related to trust transfer and the factor analysis.

Table 2. Factor Analysis

Cons-truct	Measurement Item		Factor Loading	Eigen Value	% of Var.	Cronbach's α
Site Relia-bility	SR1	This website is reliable.	0.743	1.074	15.343	0.857
	SR2	I really trust this website.	0.86			
	SR3	I do not feel that I need to be careful in using this website	0.797			
	SR4	This website satisfies my expectation	0.746			
product Infor-mation Relia-bility	SC1	The product information of this website gives me trust.	0.938	4.682	66.884	0.975
	SC2	The product information of this website is relatively reliable.	0.912			
	SC3	The product information of this website is reliable in overall.	0.923			

As suggested in Table 2, the factor loadings and statistical results on the level of trust for the questionnaire items on website reliability and product information reliability exceed the given standards, so we can conclude that the validity and reliability were supported with statistical significance. Next, to measure the match rate between the cognitive decision-making criteria and the physiological response, we asked subjects to choose three criteria they use in their decision-making in the case of both high and low involvement products. We considered the three criteria each subject reported to be his or her cognitive decision-making criteria. In addition, through the experiment using Eye-Tracker, we set an AOI for each section to use as the purchase decision-making criteria (Fig. 1). We measured the fixation length for each AOI and regarded the three criteria with the highest fixation length as the result originating from the physiological reaction.

We compared the three criteria from the cognitive decision-making process and the physiological reaction to derive the match rate for each subject. Table 3 presents the matrix of the finalized average match rate.

Table 3. The match rate depending on the degree of trust transfer and product involvement

		Product Involvement	
		High Involvement (N=70, MR*=61.0%))	Low Involvement (N=70, MR*=44.8%)
Degree of trust transfer	High (N=98, MR*=52.7%))	61.7%	44.4%
	Low (N=42, MR*=53.2%)	59.4%	45.6%

* MR : Match rate

5.2 Research Results

We analyzed the descriptive statistics using SPSS 12.0 and applied the following methods according to the research questions and hypotheses. First, we carried out average comparison and independent sample t-tests to investigate if there is a difference in the match rate between the cognitive decision-making criteria and physiological characteristics depending on the level of product involvement and degree of trust transfer (Table 4).

Table 4. Independent Samples t-Test

Classification		N	Mean	*t*	Sig.	Mean Difference
Involvement	High	70	61,0%	4.229	0.000***	0.162
	Low	70	44.8%			
Degree of trust transfer	High	98	52.7%	-0.102	0.919	-0.005
	Low	42	53.2%			

*** P < 0.001

The average match rate between the cognitive decision-making criteria and physiological reaction in the high involvement product purchase group was 61% and that of the low involvement product purchase group was 44.8%. The difference between the two average match rates was statistically significant. However, the average match rates of the high and low trust transfer groups were 52.7% and 53.2%, respectively. The match rate of the low trust transfer group was higher than that of the high trust transfer group but the difference was not statistically significant. Therefore, Hypothesis 1 is accepted but Hypothesis 2 is rejected.

In the case of high involvement, the match rate between the cognitive decision-making criteria and the physiological reaction is naturally high as the customer carries out a more active information search, while the match rate in the case of the low involvement group, in which the customer carries out a limited information search, is low. However, the difference in the match rate depending on the degree of trust transfer was not statistically significant. This may imply that trust in a website itself does not transfer to trust in product information on the website, and further research on this topic should be carried out in the future.

To judge the statistical significance of the difference in the match rate between the cognitive decision-making criteria and physiological reaction in the case of the high and low involvement product purchase depending on the degree of the trust transfer, t-tests and Mann-Whitney U tests were carried out. The Mann-Whitney U test is a nonparametric test that ranks the data, and can be used even when the size of the sample is small.

As described in Tables 5 and 6, regardless of the degree of trust transfer, the match rate in the case of purchasing the high and low involvement products was statistically significant. This shows that the difference in the match rate depending on the level of product involvement remains the same when the group is divided according to the degree of trust transfer. Therefore, Hypotheses 1-1 and 1-2 are accepted.

Table 5. Independent Samples t-Test

Degree of trust transfer	Product Involvement	N	Mean	*t*	Sig.	Mean Difference
High	High	47	61.7%	3.904	0.000***	-0.012
	Low	51	44.4%			

*** P < 0.001

Table 6. Mann-Whitney U Test

Degree of trust transfer	Product Involvement	N	Mean Rank	Mean	Std. Deviation
Low	High	23	24.717	59.4%	0.283
	Low	19	17.605	45.6%	0.199
Mann-Whitney U=144.5					**P=0.043***

* P < 0.05

We conducted t-tests to make a judgment on whether the match rate differed for the high and low trust transfer groups depending on the product involvement. As you can see in Table 7, regardless of product involvement, the difference in the match rate between the two groups was not statistically significant. This shows that the difference in the match rate depending on the degree of trust transfer remains the same when dividing the groups based on the product involvement level. Therefore, Hypotheses 2-1 and 2-2 are rejected.

Table 7. Independent Samples t-Test

Product Involvement	Degree of trust transfer	N	Mean	*t*	Sig.	Mean Difference
High	High	47	61.7%	0.351	0.726	0.023
	Low	23	59.4%			
Low	High	51	44.4%	-0.221	0.826	-0.012
	Low	19	45.6%			

6 Discussion and Conclusion

The current study shows that the difference in the match rate between cognitive thinking and the physiological reaction depending on the degree of trust transfer was not statistically significant. This means that the degree of trust transfer does not have a significant effect in the relationship between the cognitive decision-making criteria and the physiological reaction, which is a new discovery. However, for this discovery to be generalized, further research should be carried out with different samples.

On the other hand, the match rate between the decision criteria in terms of cognitive thinking and the ranks of the fixation length in the physiological reaction was higher when the purchase decision was made for a high involvement product. Customers have a tendency to search for more information in the case of high involvement products and have a tendency to become more careful in making decisions, so we can conclude that the physiological reaction is more likely to match the cognitive process when there is high involvement. However, in the case of low involvement products, participants did not put much effort in the information search, and there seemed to be much space for stereotyped concepts to be used in their decision-making, so we can infer that the probability that the physiological reaction differs from the cognitive criteria is relatively high.

According to the present results, the level of involvement is associated with a difference in the level of information that a customer requires. This also affects the probability of actual consumption taking place depending on how the product is presented to the customer. Therefore, the amount of information provided to the customers should be based on the level of involvement for the product.

This research has limitations due to the relatively small size of the sample and the fact that the experiment was carried out in a laboratory, so subjects might not behave as they would in a "real world" environment. In addition, the subjects who joined in the experiment were composed of university students who were a similar age, so there is difficulty in generalizing the results. In future research, we will apply other methods and conduct the experiment with a different subject group in order to be able to generalize the results.

Acknowledgment. This research was supported by WCU(World Class University) program through the National Research Foundation of Korea funded by the Ministry of Education, Science and Technology (Grant No. R31-2008-000-10062-0).

References

1. Cacioppo, J.T., Petty, R.E.: Physiological responses and advertising effects: Is the cup half full or half empty? Psychology and Marketing 2(2), 115–126 (1985)
2. Cho, Y., Im, I., Fjermestad, J., Hiltz, S.: The Impact of Product Category on Customer Dissatisfaction Cyberspace. Business Process Management Journal 9(5), 635–651 (2003)
3. Corritorea, C.L., Krachera, B., Wiedenbeck, S.: On-line trust: concepts, evolving themes, a model. International Journal of Human-Computer Studies 58(6), 737–758 (2003)

4. Hammond, K.R.: Human Judgment and Social Policy: Irreducible Uncertainty, Inevitable Error, Unavoidable Injustice. Oxford University Press, New York (1996)
5. Hastie, R., Dawes, R.M.: Rational choice in an uncertain world. SAGE Publications, Thousand Oaks (2001)
6. Heider, F.: The psychology of interpersonal relations. Wiley, New York (1958)
7. Kahneman, D.: Attention and effort. Prentice-Hall, Englewood Cliffs (1973)
8. Lee, K.C., Chung, N.: Cognitive Map-based Web Site Design: Empirical Analysis Approach. Online Information Review 30(2), 139–154 (2006)
9. Robertson, J.H., Schefer, R.: E-loyalty: Your Secret Weapons on the Web. Harvard Business Review 7(8), 105–113 (2000)
10. Rothschild, M.M., Gaidis, W.C.: Behavioral learning theory: It's relevance to marketing and promotion. Journal of Marketing 45(2), 70–78 (1981)
11. Stewart, K.J.: Trust transfer on the World Wide Web. Organization Science 14(1), 5–17 (2003)
12. Tan, Y.-H., Thoen, W.: Toward a generic model of trust for electronic commerce. International Journal of Electronic Commerce 5(2), 61–74 (2000-2001)
13. Tversky, A., Kahneman, D.: Judgment under uncertainty: Heuristics and biases. Science 185(4157), 1124–1131 (1974)

A Comparison of Buying Decision Patterns by Product Involvement: An Eye-Tracking Approach

Do Young Choi[1], Min Hee Hahn[2], and Kun Chang Lee[3,*]

[1] Solution Business Unit, LG CNS Co., Ltd,
Seoul 100-725, Republic of Korea
[2] SKK Business School, Sungkyunkwan University,
Seoul 110-745, Republic of Korea
[3] SKK Business School, WCU Professor at Department of Interaction Science,
Sungkyunkwan University,
Seoul 110-745, Republic of Korea
{dychoi96,minheehahn,kunchanglee}@gmail.com

Abstract. This research investigated whether buying the decision process differs by the level of product involvement. We analyzed visual attention based on the eye-tracking technique to explore the cognitive characteristics of buying decisions. More specifically, we observed visual attention of involvement by conducting experiments in a website environment. Through eye-tracking experiments, we applied physiological data in order to test our research hypotheses regarding the buying decision process and product involvement, measuring fixation length as visual attention. The results of the eye-tracking experiment showed that the decision process for high involvement products is more complicated than that for low involvement products.

Keywords: Product Involvement, Buying Decision, Eye-Tracking.

1 Introduction

Individuals' decision-making mechanisms contain inconsistencies and errors, and individuals tend to depend on a few specific pieces of information or cues when processing information [7]. Furthermore, all information is not considered when making a decision even though much information is available. Rather, each individual applies his or her own efficient decision rule to specific information [14].

Involvement is a useful concept for exploring whether individuals use different cognitive tools during information search and decision-making processes. Individuals in a high involvement situation are highly motivated to gather as much information as possible and to pay more attention to the purchase, and have a tendency to utilize many cognitive resources. On the other hand, individuals in a low involvement

* Corresponding Author.

J.-S. Pan, S.-M. Chen, N.T. Nguyen (Eds.): ACIIDS 2012, Part III, LNAI 7198, pp. 37–46, 2012.

situation tend to allocate fewer cognitive resources to the decision-making process because they have low motivation related to the information search and attention to the purchase. Therefore, it is generally accepted that every customer does not consider all of the available information when making a buying decision. Rather, they may use specific pieces of information related to the buying situation. The high involvement situation requires focusing more attention on the information search in a buying decision, while the low involvement situation requires less attention from customers.

The eye-tracking method was recently used to measure an individual's visual attention. Eye-tracking is a physiological technique used to sense visual attention by tracing eyesight, and has recently been adopted in various areas such as the usability and psychological analysis of customers in marketing research. In the current research, we investigated whether buying decisions differ for high and low involvement products by analyzing visual attention (through the measurement of eye movements using the eye-tracking technique) regarding the cognitive characteristics of a buying decision. More specifically, we observed visual attention of involvement by conducting experiments in an online environment. In order to analyze visual attention, we used fixation length to measure eyeball fixation and movement path items, which the eye-tracking technique provides.

2 Theoretical Foundation and Hypotheses

2.1 Product Involvement

Involvement refers to special attention paid to an important object on the basis of an individual's unique desires, values, and interests [15]. Product involvement can also be defined as an individual's perceived level of importance or interest [1, 5]. Meanwhile, several research studies on the effects of product involvement in customers' buying decision processes have been conducted. According to the research of Engel *et al.* [5], customers' information searches become more active and their alternative assessment becomes more complex when the level of involvement increases. Rust *et al.* [12] argued that value equity, which is determined by product quality, price, and convenience, is relatively important in the buying decision for high involvement products. On the other hand, brand equity becomes relatively more important in purchase decisions for low involvement products.

According to the explanation of customer attitude formation in the elaboration likelihood model, in the case of high product involvement, intrinsic clues are considered more important than extrinsic clues when assessing a product because customers tend to pay more attention to the intrinsic attributes of products. However, in the case of low product involvement, extrinsic clues are considered more important factors in product assessment [9].

2.2 Buying Decision

It is generally accepted that a reasonable purchase decision-making process consists of five stages: problem identification, information search, alternative assessment,

product selection, and response after purchase [5]. Problem identification can be considered as the stage when a distinct desire for the product is perceived, and occurs when customers perceive a distinct difference between the actual and desired states [8, 13]. Although the information search stage involves a conscious effort to find products that satisfy the customer's needs, its purpose is to gain pleasurable value, such as a change in surroundings and enjoyment. In other words, customers do not always search for products in order to make a purchase, and may research products with a continued interest even though they do not have an intention to buy [2]. The alternative assessment stage involves evaluating which products can satisfy a customer's needs after narrowing the choices down to several alternatives based on the information search. During reasonable decision making processes, customers evaluate product attributes based on several criteria and methods, and try to cognitively select the best alternative.

Goldsmith and Horowitz [6] found that the information search processes of customers intending to make a purchase through the Internet become more complex because customers with no space limitations can access many alternative products and a huge amount of information. As a result, more time and expenses are required to make a reasonable buying decision. According to Goldsmith and Horowitz, customers consume cognitive and emotional resources in this process.

2.3 Research Hypotheses

According to Rothschild and Gaidis [11], customers' level of activity in gathering information increases and their evaluation processes become more complex when they consider purchasing a high involvement product because they tend to use all of the given information for their buying decision. Moreover, Robertson *et al.* [10] stated that the level of involvement could change customer behaviors such as purchase-related information searching and processing.

Research on the relationship between involvement and the characteristics provided by an online environment has shown that highly involved customers show a positive attitude toward online shopping. For example, Cho *et al.* [4] found that the degree of a customer's information search is related with the degree of involvement. Based on this finding, we can argue that customers considering low involvement products do not put much effort into the information search. On the other hand, customers considering high involvement products conduct more detailed information searches, exerting more effort. In this sense, when the degree of product involvement is high, customers consider many factors in their buying decision processes and consequently require a large amount of information in order to find differences among the products. On the other hand, in the case of low involvement products, customers do not tend to search for information about differences among the products and consequently process related information through alternative channels [3].

Based on the previous studies discussed above, we developed a research question and related hypotheses regarding the relationship between customers' buying decisions and the level of product involvement. Our research question is whether the decision processes regarding the purchase of high involvement products are more

complicated than those used for the purchase of low involvement products. Our research hypotheses are as follows.

H1. The buying decision process for high involvement products takes more time than the buying decision process for low involvement products.

H2. The buying decision process for high involvement products includes more consideration factors than the buying decision for low involvement products.

3 Experiment and Analysis

3.1 Experimental Design and Procedure

In order to explore the relationship between the buying decision and involvement processes for high and low involvement products, an experiment was conducted with 70 university students in Korea in September 2011. Each student participated in two experiments related to buying decisions with a high involvement product and a low involvement product on an Internet website. The high and low involvement products were a used car and used book, respectively. Participants were asked to make a decision based on using a specialized mediating site for used products. The length of the experiment was about 20 minutes. Information on demographics and cognitive variables was collected through a questionnaire after the experiment ended. As summarized in Table 1, participants were asked to choose one seller and a selling condition after considering several seller conditions, such as product quality, price, credit, and so on, with the target products fixed (used car and used book). At this point, participants selected one seller on the basis of the information provided about the products on the website. In the experiment, participants were provided with seven pieces of information for each high and low involvement product, such as seller, product information, credit, price, register date, and quality in order to match the real environment.

In order to measure eye movement and gather related data during the buying decision process on the website, Eye-Tracker, which was developed by Tobii Technology Corporation, was used. Eye-Tracker can measure participants' visual attention, including eyeball fixation and saccade. Eyeball fixation shows how long a participant's eyes stay focused on a certain area and saccade is the momentary movement between eyeball fixations. In this research, visual attention was based on the fixation length, or how long the participant's eyes stayed focused on certain product information displayed on the website.

The experimental procedure was as follows. First, a calibration test was conducted in order to correctly trace a participant's eye movement before starting the product buying experiment. Second, four buying processes were displayed sequentially to participants, as shown in Figure 1, following the process of a real purchase in the online shopping environment. The sequence of the buying process is as follows: (1) access to the specialized website for used products, (2) search for product, (3) evaluation and buying decision, and (4) confirmation of the selected product. Lastly, participants were asked to complete the questionnaire.

Table 1. Experiments

Experiment	Experiment for Buying Decision
Experiment 1 (High Involvement Product)	• Buying Decision of the used-car by the process of online shopping through on-line dealing site for used cars. • Traced participant's eye movement and decision time with eye-tracking equipment during the whole processes of used-car shopping. • Consideration factors for buying decision: Seller Information, Product Information, Credit, Price, Register Date, Quality1(Release Date), Quality2(Mileage)
Experiment 2 (Low Involvement Product)	• Buying Decision of the used-book by the process of online shopping through on-line dealing site for used books. • Traced participant's eye movement and decision time with eye-tracking equipment during the whole processes of used-book shopping. • Consideration factors for buying decision: Seller Information, Product Information, Credit, Price, Register Date, Quality1(Condition), Quality2(Delivery Quality)

(a) Experiment of Buying Processes for High Involvement Product(Used-car)

(b) Experiment of Buying Processes for Low Involvement Product(Used-book)

Fig. 1. Experiment of Buying Processes

3.2 Experimental Results and Analysis

We analyzed the data gathered during the third process (the evaluation and buying decision process) in order to compare the complexity of the buying decision for high and low involvement products from the perspective of the decision time required to make the purchase and where the user's visual attention was focused. Among the data gathered using Eye-Tracker, we used two variables for analysis: decision time required and fixation length, as defined by the area of interest (AOI). Seven AOI areas were set by the analyzing software Tobii provided in accordance with seven factors in the buying decision: seller, product information, credit, price, register date,

quality item 1, and quality item 2. In order to verify Hypotheses 1 and 2, the complexities of the buying process for high and low involvement products were analyzed using a t-test in SPSS 13.0.

(1) Analysis of Decision Time Required

Table 2 shows the t-test results regarding Hypothesis 1. The results show a significant difference between high and low involvement products in terms of the time required to make a buying decision. The t-value of the decision time required is 2.599, showing a significant difference between the buying decision related to high and low involvement products. The mean decision times required for the high and low involvement products are 32.136 and 25.978 seconds, respectively. Therefore, Hypothesis 1 was accepted, as the purchase of a high involvement product requires a longer decision time than the purchase of a low involvement product, as shown in Table 2.

Table 2. T-test Results for Decision Time Comparison

Decision Time	Mean		Standard Deviation		t-value	Sig.	Mean Difference
	HIP (n=70)	LIP (n=70)	HIP (n=70)	LIP (n=70)			
Decision Time	32.136	25.978	15.567	12.268	2.599	0.010*	6.156

*$p<0.05$

HIP: High Involvement Product, LIP: Low Involvement Product

(a) Heat Map of High Involvement Product (b) Heat Map of Low Involvement Product

Fig. 2. Heat Map measured by Eye-Tracking

(2) Analysis of Consideration Factors for Buying Decision

In this research we analyzed how many factors were considered during the buying decision process for high and low involvement products based on the perspective of visual attention. We considered visual attention to be how long the individual's eye focused on the consideration factors on the website provided during the buying decision. When the fixation length of a consideration factor is long, the factor can be regarded as a more important factor in the buying decision.

Figure 3 presents the heat map, which is a visualization tool provided by Tobii Eye-Tracker. It shows the average fixation length of participants, which can be interpreted as the degree of visual attention for each consideration factor, namely AOI. As seen in Figure 2, the purchase of a high involvement product leads a user to consider more factors than when purchasing a low involvement product.

In order to statistically verify the intuitive results, the data on fixation length for each AOI provided by Eye-Tracker were analyzed. Figure 3 and Table 3 show the comparison of means among the seven factors and the t-test results to verify whether the buying decision differs for the purchase of high and low involvement products.

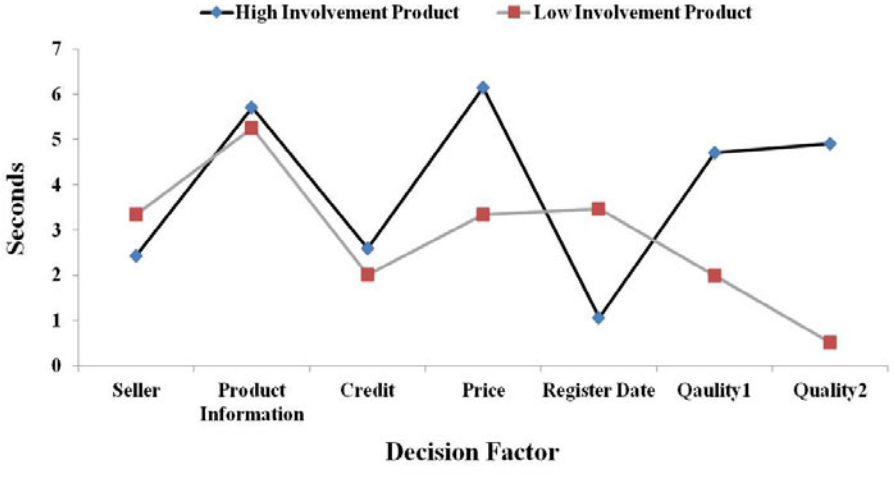


Fig. 3. Comparison of Mean-value of Consideration Factors for Buying Decision

The t-values for product information and credit were 0.698 and 1.747, respectively, showing that there were no differences between the two product involvement types in terms of fixation length. However, five factors (seller, price, register date, quality 1, and quality 2) showed significant differences in terms of fixation length. Four factors (seller, price, quality 1, and quality 2) had a longer fixation length for the high involvement product than for the low involvement product. Moreover, four factors (product information, price, quality 1, and quality 2) had a fixation length of more than 4.000 seconds for the high involvement product. This indicates that participants considered these four factors as the most important criteria in the buying decision during the experiment. On the other hand, only one factor, product information, had a

fixation length of more than 4.000 seconds in the case of the low involvement product. Therefore, we can infer that the factors considered are more important for high involvement products than for low involvement products in terms of customers' buying decisions.

Table 3. T-test Results for Consideration Factors Comparison

Consideration Factor	Mean		Standard Deviation		t-value	Sig.	Mean Difference
	HIP (n=70)	LIP (n=70)	HIP (n=70)	LIP (n=70)			
Seller	2.431	3.344	1.709	2.429	-2.570	0.011*	-0.912
Product Information	5.718	5.254	4.819	2.758	0.698	0.486	0.463
Credit	2.586	2.001	2.119	1.833	1.747	0.083	0.585
Price	6.144	3.330	3.151	2.235	6.095	0.000***	2.814
Register Date	1.064	3.466	1.615	2.908	-6.042	0.000***	-2.402
Quality1	4.713	1.987	4.262	1.932	4.874	0.000***	2.726
Quality2	4.910	0.515	3.979	0.618	9.133	0.000***	4.396

*p<0.05, **p<0.01, ***p<0.001

HIP: High Involvement Product, LIP: Low Involvement Product

Table 4. T-test Results for Consideration Factors Comparison by Fixation Length

Fixation Length	Mean		Standard Deviation		t-value	Sig.	Mean Difference
	HIP (n=70)	LIP (n=70)	HIP (n=70)	LIP (n=70)			
More than 8.0 seconds	0.87	0.29	1.034	0.640	4.028	0.000***	0.586
More than 7.0 seconds	1.11	0.54	1.210	0.943	3.116	0.002**	0.571
More than 6.0 seconds	1.54	0.87	1.401	1.076	3.181	0.002**	0.671
More than 5.0 seconds	2.03	1.37	1.560	1.194	2.799	0.006**	0.657
More than 4.0 seconds	2.60	1.90	1.663	1.385	2.706	0.008**	0.700
More than 3.0 seconds	3.49	2.66	1.576	1.710	2.981	0.003**	0.829
More than 2.0 seconds	4.43	3.60	1.538	1.536	3.190	0.002**	0.829
More than 1.0 second	5.69	5.06	1.015	1.318	3.162	0.002**	0.629

*p<0.05, **p<0.01, ***p<0.001

HIP: High Involvement Product, LIP: Low Involvement Product

A t-test was also conducted to verify whether there were statistically significant differences in the number of consideration factors in accordance with the range of fixation length. Table 4 shows the t-test results when the fixation length ranges were set every few seconds. The number of consideration factors used for the high involvement product is higher than that used for the low involvement product, and the difference is statistically significant. Therefore, the results support Hypothesis 2.

4 Discussion and Concluding Remarks

The experiment investigating the customer buying decision with an eye-tracking technique showed that there are statistically significant differences in eye movements during the buying decision between high and low involvement products. Specifically, the results for fixation length, regarded as visual attention in this research, showed that the decision process for high involvement products is more complicated than that for low involvement products. First, there is a significant difference in the time required for a buying decision with high and low involvement products. The decision time required when purchasing a high involvement product is longer than that when purchasing a low involvement product. Second, the eye-tracking technique demonstrated that there is a significant difference in the number of factors considered in the buying process for high and low involvement products. Five of the seven factors considered in the experiment – seller, price, register date, quality 1, and quality 2 – differed significantly for the high and low involvement products in terms of fixation length. The fixation length for the high involvement product was longer than that for the low involvement product for seller, price, quality 1, and quality 2. Therefore, we can infer that more factors are considered when purchasing a high involvement product as compared with the purchase of a low involvement product.

In this research, the eye-tracking technique was adopted in order to reconfirm previous studies on differences in the buying process depending on the level of product involvement. By measuring fixation length as visual attention using Eye-Tracker, we applied physiological data through experiments in order to verify our research hypotheses regarding the buying decision processes at different levels of product involvement. Nevertheless, there are several limitations of this research. First, the experiments were conducted only on an online website, which means that there is difficulty in generalizing the results to the broader population. Second, the sample size was relatively small and the participants consisted only of university students. In future studies, more consideration factors, such as an offline situation, should be applied in order to generalize the research findings.

Acknowledgment. This research was supported by WCU(World Class University) program through the National Research Foundation of Korea funded by the Ministry of Education, Science and Technology (Grant No. R31-2008-000-10062-0).

References

1. Antil, J.H.: Conceptualization and operationalization of involvement. Advances in Consumer Research 13, 207–210 (1984)

2. Bloch, P.H., Sherrell, D.L., Ridgway, N.M.: Consumer search: An extended framework. Journal of Consumer Research 13(1), 119–126 (1986)
3. Cacioppo, J.T., Petty, R.E.: Physiological responses and advertising effects: Is the cup half full or half empty? Psychology and Marketing 2(2), 115–126 (1985)
4. Cho, Y., Im, I., Fjermestad, J., Hiltz, S.: The Impact of Product Category on Customer Dissatisfaction Cyberspace. Business Process Management Journal 9(5), 635–651 (2003)
5. Engel, J.F., Blackwell, R.D., Miniard, P.W.: Consumer Behavior, 8th edn., pp. 44–48. The Dryden Press, New York (1995)
6. Goldsmith, R.E., Horowitz, D.: Measuring motivations for online opinion seeking. Journal of Interactive Advertising 6(2), 1–16 (2006)
7. Hastie, R., Dawes, R.M.: Rational choice in an uncertain world. SAGE Publications, Thousand Oaks (2001)
8. Hawkins, D.I., Best, R., Coney, K.A.: Consumer behavior: Building marketing strategy, 9th edn. McGraw Hill, New York (2003)
9. Petty, R.E., Cacioppo, J.T., Schmann, D.: Central and peripheral routes to advertising effectiveness: the moderating role of involvement. Journal of Consumer Research 10, 135–145 (1983)
10. Robertson, J.H., Schefer, R.: E-loyalty: Your Secret Weapons on the Web. Harvard Business Review 7(8), 105–113 (2000)
11. Rothschild, M.M., Gaidis, W.C.: Behavioral learning theory: It's relevance to marketing and promotion. Journal of Marketing 45(2), 70–78 (1981)
12. Rust, R.T., Zeithaml, V.A., Lemon, K.N.: Driving Customer Equity. Free Press (2000)
13. Solomon, M.R.: Consumer behavior: Buying, having and being, 5th edn. Prentice Hall, Upper Saddle River (2002)
14. Tversky, A., Kahneman, D.: Judgment under uncertainty: Heuristics and biases. Science 185(4157), 1124–1131 (1974)
15. Zaichkowsky, J.L.: Familiarity: product use, involvement or expertise? Advances in Consumer Research 12, 296–299 (1985)

The Influence of Team-Member Exchange on Self-reported Creativity in the Korean Information and Communication Technology (ICT) Industry

Dae Sung Lee[1], Kun Chang Lee[2,*], and Nam Yong Jo[3]

[1] SKK Business School, Sungkyunkwan University,
Seoul 110-745, Republic of Korea
leeds1122@gmail.com
[2] SKK Business School
WCU Professor at Department of Interaction Science
Sungkyunkwan University, Seoul110-745, Republic of Korea
kunchanglee@gmail.com
[3] SKK Business School, Sungkyunkwan University,
Seoul 110-745, Republic of Korea
higlobe@naver.com

Abstract. There have been some studies on the relationship between social exchange relationships and creativity in organizations, but researchers have little explored the effect of TMX (team-member exchange) and coworker behaviors on creativity. In this respect, we introduced into our research model TMX and coworker behaviors as social or contextual factors which might influence individual creativity within Korean ICT companies. Although CHS (Coworker helping and support) does not have a sufficient mediating effect on individual creativity, that factor is positively related to Individual creativity. On the other hand, FC (Feedback from Coworkers) does not influence individual creativity at all. However, we found that TMX directly and strongly influences individual creativity. Therefore, organizations should encourage employees to build a high level of social exchange relationships with coworkers.

Keywords: Individual creativity, TMX (Team-Member Exchange), Coworker helping and support, Feedback from coworkers.

1 Introduction

In the global competitive business environment, employee creativity has been regarded as the building block for the innovation that enhances an organization's adaptability and growth [22]. Like telephones, computers, television, radio and other ubiquitous communication devices, Information and Communication Technologies (ICTs) are fast becoming essential components of our day-to-day lives. Therefore, it is necessary to investigate employee creativity in the ICT industry, as the industry requires a high level of organizational creativity.

* Corresponding author.

J.-S. Pan, S.-M. Chen, N.T. Nguyen (Eds.): ACIIDS 2012, Part III, LNAI 7198, pp. 47–54, 2012.

Research by Amabile and her associates documents the value of examining the creativity of individuals and groups within their relevant social settings [11]. *The interactionist model of creativity* emphasizes social influences from the group and contextual influences from the organization, both of which have effects on individual and team creativity [23, 24]. In this respect, our model selected team-member exchange and coworker behaviors (Help, Support and Feedback) as social or contextual factors that might influence individual creativity within an organization. Although some studies have examined the relationship between LMX (Leader-Member Exchange) and employee creativity, few researchers have examined the effect of TMX (Team-Member Exchange) on creativity. With this study, we hope to contribute to the literature on the interaction between social exchange relationships and creativity.

2 Creativity and Its Antecedents within Organization

Creativity is generally defined as the ability to produce work that is both novel and useful [21]. If this concept is used in the context of an organization, organizational creativity is the production of a valuable, useful new idea, procedure, product, service, or process by individuals working together in a complex social system [23]. In order to understand organizational creativity, researchers need to investigate the creative product, process, person and situation. Above all, it is important to understand the way in which each of these components interacts with the others [2, 8]. Based on the *interactionist model of creativity*, we emphasized social settings within an organization, and selected TMX (Team-Member Exchange), CHS (Coworker Helping and Support) and FC (Feedback from Coworkers) as social and contextual factors that might influence individual creativity within an organization.

TMX (Team-Member Exchange), as introduced by Seers [17], is generally defined as a way to assess the reciprocity between a member and his or her team [13]. Therefore, TMX represents the quality of team members' working relationships with their peer group [17]. The qualities to measure are those related to a member's willingness to assist other team members, share feedback, and contribute ideas [17]. Seers, Petty, and Cashman [18] found measures of TMX to be higher in autonomous teams than in traditional work groups. They concluded that the more a group was self-managed, the greater the need for members to engage in supportive reciprocal exchanges with one another.

As individual creativity is often exhibited in a team context [19], it is important to understand, theoretically and empirically, how a member's social exchange relationships with the other members influence the member's creativity. High-quality TMX may allow a team member to interact sufficiently with the other members, heeding their work behaviors and expressed ways of thinking [12]. Therefore, TMX might be considered as a social factor that influences creativity. Thus:

Hypothesis 1: TMX will contribute positively to Coworker helping and support.
Hypothesis 2: TMX will contribute positively to Feedback from coworkers.
Hypothesis 3: TMX will contribute positively to Individual creativity.

CHS (Coworker Helping and Support) is defined as coworkers helping an employee with his or her tasks by sharing knowledge and expertise or providing encouragement and support. If coworkers are helpful and supportive, it would be easy for employees to use coworkers as a stepping stone to new ideas [6, 16]. Thus:

Hypothesis 4: CHS will contribute positively to Individual creativity.

If FC (Feedback from Coworkers) is useful, the feedback is valuable information provided by coworkers that enables an employee to make improvements on the job. When the feedback is related to job improvement, employees are likely to pay attention to learning and making improvements on the job. In this situation, they may be stimulated to have different perspectives and come up with new and useful ways of performing tasks [6, 14, 16]. Thus:

Hypothesis 5: FC will contribute positively to Individual creativity.

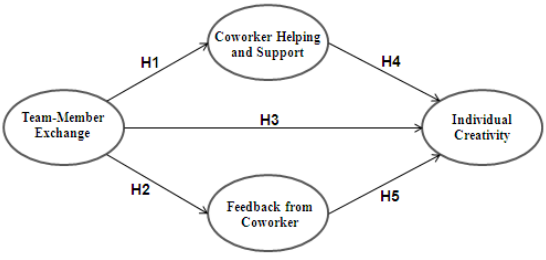


Fig. 1. Research model

3 Research Methodology

3.1 Data Collection

The participants were employees in Korean ICT companies. We collected a total of 365 surveys from a company that administered them to the participant employees. To avoid selection bias, inconsistent or incomplete responses were eliminated from the dataset. Moreover, all respondents above senior managers were excluded because our questionnaire was for employees who interacted with each other horizontally. This process yielded 320 useable cases (87.7 percent). Table 1 reports the sample characteristics.

Table 1. Sample Characteristics

Characteristics		Frequency	Percent
Gender	Male	259	80.9
	Female	61	19.1
Age	19 – 29	92	28.8
	30 – 39	185	57.8
	40 – 49	42	13.1
	50 – 59	1	0.3
Position	Assistant	63	19.7
	Senior Assistant	100	31.3
	Manager	86	26.9
	Senior Manager	71	22.2
Total		320	100

3.2 Measures

We measured each item in our model on a seven-point Likert scale, with answers ranging from one ‘‘strongly disagree” to seven ‘‘strongly agree” (see Table 2). The items in the survey were developed by adapting existing measures validated by previous studies or by converting construct definitions into a questionnaire format.

Table 2. Construct and Measurement

Construct	Items	Measurement	Related literature
Team-member exchange	TMX1	I am flexible about switching jobs with others in my work group.	Seers (1989) [17]
	TMX2	Other group members recognize my potential.	
	TMX3	Other group members understand my problems.	
	TMX4	I am wiling to finish work assigned to others.	
	TMX5	Others are willing to finish work assigned to me.	
Coworker helping and support	CHS1	My coworkers willingly share their expertise with each other.	Podsakoff, Ahearne, and MacKenzie (1997) [15]
	CHS2	My coworkers help each other out if someone falls behind in his/her work.	
	CHS3	My coworkers encourage each other when someone is down.	
	CHS4	My coworkers try to act like peacemakers when there are disagreements.	
Feedback from coworkers	FC1	I find the feedback I receive from my coworkers very useful.	Zhou and George (2001) [25]
	FC2	My coworkers provide me with valuable information about how to improve my job performance.	
	FC3	The feedback I receive from my coworkers helps me improve my job performance.	
Individual creativity	IC1	I am willing to propose a new idea or method ahead of my coworkers.	Ettlie O'Keefe (1982) [4], Scott and Bruce (1994) [16], Zhou and George (2001) [25]
	IC2	I generally use existing methods and instruments as new ones.	
	IC3	I employ a proper planning and scheduling in order to implement a new idea.	
	IC4	I propose a new and better method to achieve a goal.	

4 Results

4.1 Measurement

Partial least squares (PLS) analysis has been used widely in theory testing and confirmation, and is appropriate for determining whether relationships such as those in which we were interested exist [7]. We used SmartPLS 2.0 to analyze the measurement and structural models. As shown in Table 3, our composite reliability

values ranged from 0.896 to 0.952, and our variance extracted by the measures ranged from 0.633 to 0.868. All of these figures are above acceptable levels.

Table 3. Confirmatory Factor Analysis

Construct	Items	Factor Loading	t-value	Cronbach's α	Composite Reliability	AVE
Team-member exchange	TMX1	0.720	21.394	0.855	0.896	0.633
	TMX2	0.805	34.930			
	TMX3	0.840	48.034			
	TMX4	0.824	38.613			
	TMX5	0.783	23.296			
Coworker helping and support	CHS1	0.825	34.032	0.860	0.906	0.707
	CHS2	0.878	49.983			
	CHS3	0.892	75.937			
	CHS4	0.763	18.469			
Feedback from coworkers	FC1	0.911	70.387	0.924	0.952	0.868
	FC2	0.943	134.994			
	FC3	0.942	115.892			
Individual creativity	IC1	0.862	47.208	0.874	0.914	0.727
	IC2	0.756	16.014			
	IC3	0.889	66.602			
	IC4	0.896	82.090			

The discriminant validity was assessed by examining the correlations among variables [7]. For satisfactory discriminant validity, the AVE (Average Variance Extracted) from the construct should be greater than the variance shared between the construct and other constructs in the model [3]. Table 4 lists the correlation matrix with the correlations among the constructs and the square root of AVE on the diagonal.

Table 4. Correlation of Latent Variables

Construct	Coworker helping and support	Feedback from coworkers	Team-member exchange	Individual Creativity
Team-member exchange	0.795			
Coworker helping and support	0.679	0.841		
Feedback from coworkers	0.590	0.743	0.932	
Individual creativity	0.526	0.468	0.389	0.853

4.2 Structural Model

We tested the structural model by estimating the path coefficients and R^2 values. The R^2 and path coefficients indicate how well the data support the hypothesized model. Figure 2 and Table 5 represent the results of the test of the hypothesized structural model. All of the R^2 values in this model are above the value of 10% recommended by Falk and Miller [5]. These results support the hypothesized relationships among the constructs in our model. As shown in Table 5, all of the hypotheses are supported except H5. However, the strength of the paths varies in the coefficient levels. The findings are discussed in the section below.

Table 5. Results of the Hypothesis Tests

Hypo-theses	Path	Coefficient (β)	t-Value
H1	Team-member exchange → Coworker helping and support	0.679	19.203**
H2	Team-member exchange → Feedback from coworkers	0.590	14.223**
H3	Team-member exchange → Individual creativity	0.383	5.583**
H4	Coworker helping and support → Individual creativity	0.195	2.274*
H5	Feedback from coworkers → Individual creativity	0.018	0.203

*p < 0.05, **p < 0.01

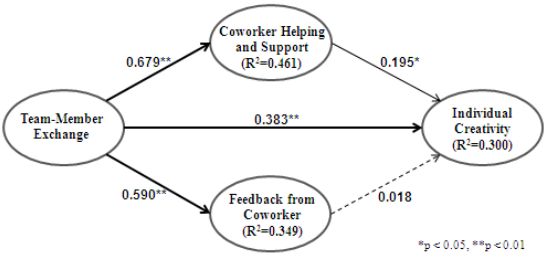


Fig. 2. Research Model Results

5 Discussion

5.1 Implications

Though CHS does not have a sufficient mediating effect on individual creativity, it does contribute positively to Individual creativity. On the other hand, FC does not influence individual creativity. Kluger and DeNisi [10] found that feedback typically has only a moderately positive influence on performance and that more than 38% of the effects are negative. This suggests that the mechanisms of feedback are not as well understood as we would like to think. For example, '*The definitions of favorable and unfavorable feedback take into account the extent to which the feedback recipient receives positive and negative feedback that upon reflection is believed to accurately reflect performance (Steelman, Levy and Snell, 2004).*'[20]. Moreover, it might be more complicated to consider the mechanism within the organizational creativity domain, in that the relationship between creativity and performance remains unclear.

From our results, we found that TMX directly and strongly influences individual creativity. As a contextual factor, the concept TMX is broader than that of coworker's behaviors (CHS and FC), in that TMX represents the intention to help teammates, share feedback, and contribute ideas [17]. Therefore, managers should pay attention to social exchange relationships such as TMX in order to enhance individual creativity within their organizations.

5.2 Limitations and Future Research

Our study targeted Korean ICT companies. Future research should pursue comparative studies in various industries. We did not consider many important antecedents of team-member exchange. Future research should focus on identifying different types of justice, team temporal scope, communication mediation, and supervisor-subordinate relationships as the antecedents of TMX (e.g., [1], [9]). Moreover, we did not fully explain the negative effects of feedback from coworkers. Future research should examine the feedback process in organizations in more depth, and examine theoretical backgrounds that may explain the negative influences we found.

Acknowledgments. This study was supported by WCU (World Class University) program through the National Research Foundation of Korea funded by the Ministry of Education, Science and Technology (Grant No. R31-2008-000-10062-0).

References

1. Alge, B.J., Wiethoff, C., Klein, H.J.: When Does the Medium Matter? Knowledge-Building Experience and Opportunities in Decision Making Teams. Organizational Behavior and Human Decision Processes 91(1), 26–37 (2003)
2. Brown, R.T.: Creativity: What Are We to Measure? In: Glover, J.A., Ronning, R.R., Reynolds, C.R. (eds.) Handbook of Creativity, pp. 3–32. Plenum Press, New York (1989)
3. Chin, W.W.: The Partial Least Squares Approach to Structural Equation Modeling. In: Marcoulides, G.A. (ed.) Modern Methods for Business Research. Lawrence Erlbaum, Mahway (1988)
4. Ettlie, J.E., O'Keefe, R.D.: Innovative Attitudes, Values, and intentions in Organizations. Journal of Management Studies 19(2), 163–182 (1982)
5. Falk, R.F., Miller, N.B.: A Premier for Soft Modeling. The University of Akron, Akron (1992)
6. Farr, J.L.: Facilitating Individual Role Innovation. In: West, M.A., Farr, J.L. (eds.) Innovation and Creativity at Work: Psychological and Organizational Strategies, pp. 207–230. John Wiley & Sons, Oxford (1990)
7. Fornell, C.R., Lacker, D.F.: Two Structural Equation Models with Unobservable Variables and Measurement Error. Journal of Marketing Research 18(1), 39–50 (1981)
8. Harrington, D.M.: The Ecology of Human Creativity: A Psychological Perspective. In: Runco, M.A., Albert, R.S. (eds.) Theories of Creativity, pp. 143–169. Sage, Newbury Park (1990)

9. Hiller, N.J., Day, D.V.: LMX and Teamwork: The Challenges and Opportunities of Diversity. In: Graen, G.B. (ed.) Dealing with Diversity: A Volume in LMX Leadership: The Series, vol. 1, pp. 29–57. Information Age Publishing, Greenwich (2003)
10. Kluger, A.N., DeNisi, A.: The Effects of Feedback Interventions on Performance: A Historical Review, a Meta-Analysis, and a Preliminary Feedback Intervention Theory. Psychological Bulletin 119(2), 254–284 (1996)
11. Lee, D.S., Seo, Y.W., Lee, K.C.: Individual and Team Differences in Self-Reported Creativity by Shared Leadership and Individual Knowledge in an e-Learning Environment. Information 14(9), 2931–2946 (2011)
12. Liao, H., Liu, D., Loi, R.: Looking at Both Sides of the Social Exchange Coin: A Social Cognitive Perspective on the Joint Effects of Relationship Quality and Differentiation on Creativity. Academy of Management Journal 53(5), 1090–1109 (2010)
13. Liu, Y., Keller, R.T., Shih, H.A.: The Impact of Team-Member Exchange, Differentiation, Team Commitment, and Knowledge Sharing on R&D Project Team Performance. R&D Management 41(3), 274–287 (2011)
14. Oldham, G.R., Cummings, A.: Employee Creativity: Personal and Contextual Factors at Work. Academy of Management Journal 39(3), 607–634 (1996)
15. Podsakoff, P.M., Ahearne, M., MacKenzie, S.B.: Organizational Citizenship Behavior and the Quantity and Quality of Work Group Performance. Journal of Applied Psychology 82(2), 262–270 (1997)
16. Scott, S.G., Bruce, R.A.: Determinants of Innovative Behavior: A Path Model of Individual Innovation in the Workplace. Academy of Management Journal 37(3), 580–607 (1994)
17. Seers, A.: Team-Member Exchange Quality: A New Construct for Role-Making Research. Organizational Behavior and Human Decision Processes 43(1), 118–135 (1989)
18. Seers, A., Petty, M.M., Cashman, J.F.: Team-Member Exchange under Team and Traditional Management: A Naturally Occurring Quasi-Experiment. Group & Organization Management 20(1), 18–38 (1995)
19. Shalley, C.E., Zhou, J., Oldham, G.R.: The Effects of Personal and Contextual Characteristics on Creativity: Where Should We Go from Here? Journal of Management 30(6), 933–958 (2004)
20. Steelman, L.A., Levy, P.E., Snell, A.F.: The Feedback Environment Scales (FES): Construct Definition, Measurement and Validation. Educational and Psychological Measurement 64(1), 165–184 (2004)
21. Sternberg, R.J.: The Nature of Creativity: Contemporary. Psychological Perspectives. Cambridge University Press, Cambridge (1988)
22. Tierney, P., Farmer, S.M., Graen, G.B.: An Examination of Leadership and Employee Creativity: The Relevance of Traits and Relationships. Personnel Psychology 52(3), 591–620 (1999)
23. Woodman, R.W., Sawyer, J.E., Griffin, R.W.: Toward a Theory of Organizational Creativity. Academy of Management Review 18(2), 293–321 (1993)
24. Woodman, R.W., Schoenfeldt, L.F.: Individual Differences in Creativity: An Interactionist Perspective. In: Glover, J.A., Ronning, R.R., Reynolds, C.R. (eds.) Handbook of Creativity, pp. 77–92. Plenum Press, New York (1989)
25. Zhou, J., George, J.M.: When Job Dissatisfaction Leads to Creativity: Encouraging the Expression of Voice. Academy of Management Journal 44(4), 682–696 (2001)

Semi-parametric Smoothing Regression Model Based on GA for Financial Time Series Forecasting

Lingzhi Wang[1,2]

[1] School of information Engineering,Wuhan University of Technology Wuhan,430070, Hubei, China
[2] Department of Mathematics and Computer, Liuzhou Teacher College, Liuzhou, 545004, Guangxi, China
wlz1974@163.com

Abstract. In this study, a novel Neural Network (NN) ensemble model using Projection Pursuit Regression (PPR) and Least Squares Support Vector Regression (LS–SVR) is developed for financial forecasting. In the process of ensemble modeling, the first stage some important economic factors are selected by the PPR technology as input feature for NN. In the second stage, the initial data set is divided into different training sets by used Bagging and Boosting technology. In the third stage, these training sets are input to the different individual NN models, and then various single NN predictors are produced based on diversity principle. In the fourth stage, the Partial Least Square (PLS) technology is used to choosing the appropriate number of neural network ensemble members. In the final stage, LS–SVR is used for ensemble of the NN to prediction purpose. For testing purposes, this study compare the new ensemble model's performance with some existing neural network ensemble approaches in terms of the Shanghai Stock Exchange index. Experimental results reveal that the predictions using the proposed approach are consistently better than those obtained using the other methods presented in this study in terms of the same measurements.

Keywords: Semi-parametric regression, Partial Least Square, Genetic Algorithm, Financial time series prediction.

1 Introduction

Effective data analysis and forecasting plays an important role in the field of financial investment. World financial markets function in a very complex and dynamic manner where high volatility and noisy data are routine. Many factors impact financial markets, including political events, general economic conditions [1,2,3]. Due to the high degrees of irregularity, dynamic manner and nonlinearity, it is extremely difficult to capture the irregularity and nonlinearity hidden in financial time series by traditional linear models such as multiple regression, exponential smoothing, autoregressive integrated moving average, etc [4,5]. Nevertheless, dealing with financial time series prediction, such

J.-S. Pan, S.-M. Chen, N.T. Nguyen (Eds.): ACIIDS 2012, Part III, LNAI 7198, pp. 55–64, 2012.

a nonlinear and complicated system, it is not adequate to predict with only one particular forecasting model.

In recent years, semi–parametric regression models have attracted a lot of research interests due to their flexibility to allow both linear and nonlinear components, as well as serially correlated errors, which enables them to better describe increasingly complex data from the real world than pure parametric or nonparametric models [6,7]. It synthesizes research across several branches of Statistics: parametric and nonparametric regression, longitudinal and spatial data analysis, mixed and hierarchical Bayesian models, Expectation-Maximization (EM) and Markov Chain Monte Carlo (MCMC) algorithms [8,9].

Semi–parametric regression is a field deeply rooted in applications and its evolution reflects the increasingly large and complex problems that are arising in science and industry. It is known that semi–parametric time series regression is often used without checking its suitability and compactness. In theory, this may result in dealing with an unnecessarily complicated model. In practice, two problems may encounter the computational difficulty caused by the spareness of the data. This is partly because the curse of dimensionality problem may still arise from using a semi–parametric time series regression model, in addition reason is difficult to determine the smoothing parameter. The choice of the smoothing parameter is crucial, when semi–parametric regression model is used to fit a smooth curve.

The financial data one of the tasks is to study the structural relationship between the present observation and the history of the data set. The problem then is to fit a high dimensional surface to a nonlinear data set. The financial market is a complex system that contains many uncertain factors, and it has been interact with all sorts of economic, political and social factors. This paper is presented a composition information and feature detection from economic factor by Partial Least Square (PLS) regression. The genetic algorithm (GA) is applied to search the optimal smoothing parameter in order to improve the smoothness of curve fitted.

The rest of this paper is organized as follows: Section 2 describes the building process of the PLS for feature detection and GA optimization smoothing parameter. Section 3 For further illustration, two real financial time series are used for testing in Section 4. Finally, some concluding remarks are drawn in Section 5.

2 The Establishment of Semi–parametric Regression Model

2.1 Extraction Composition Information by PLS Method

PLS is a new multivariate statistical data analysis method which can be used to effective dimension reduction and feature extraction for systems. It has been widely used in engineering technology, and the basic idea is as same as the principal components analysis regression modeling method [10]. However, the main difference is that, in information integration and screening process, the

PLS not only to consider the dimension reduction and information integration of independent variables, but also to consider the explanation ability of new information to dependent variables [11].

Let $X_0 = (x_{ij}, i = 1, 2, \cdots, n, j = 1, 2, \cdots, m)$ be the independent variables matrix and $y_0 = (y_{i1}, i = 1, 2, \cdots, n,)$ be the dependent variables matrix. We list the detail calculation steps as follows:

(1) Normalized the independent variables matrix and the dependent variables matrix, and obtained the normative data X_0^*, Y_0^*;

(2) Calculating spindles

$$\omega_i = \frac{X_{i-1}^* Y_{i-1}^*}{\|X_{i-1}^* Y_{i-1}^*\|}, i = 1, 2, \cdots, T \tag{1}$$

and then we can get the ith aggregate variable $F_i = X_{i-1}^* \omega_i$. We estimate F_i and X_{i-1}^* by ordinary least square regression, where the regression coefficient is $P_i = \frac{X_{i-1}^* t_i}{\|t_i\|^2}$, and calculate the residual matrix $X_i^* = X_{i-1}^* - F_i P_i$;

(3) Cross validation test: if $Q_i^2 \geq 0.0975$, then we continue the calculation; otherwise, it is completed;

(4) Exacting T components $F_1, F_2, \cdots F_T$, as composition information.

2.2 Semi–parametric Smooth Regression Model

The predictor function of the semi–parametric regression model consists of a parametric linear component and an arbitrary nonparametric component depending on a design variable t denoting time in this paper. Consider a semi–parametric regression model of the following form

$$y_i = x_i^T \beta + h(z_i) + \varepsilon_i \tag{2}$$

where y_i are responses, x_i and z_i are are called the design points, β is an unknown parameter vector representing the linear component, $h(\cdot)$ is an unknown smooth function defined on $[0, 1]$ for the nonlinear component, ε_i are unobservable random errors, zero mean random variables with a common variance σ^2 [12,13].

The corresponding sum of squares equation is

$$SS(h, \beta) = \sum_{i=1}^{n} (y_i - x_i^T \beta - h(z_i))^2 + \lambda \int_a^b [h''(z_i)]^2 dt \tag{3}$$

A solution can be obtained by minimizing equation (3), and denotes the roughness penalty of a cubic smoothing spline with knots at $t_1, t_2, \cdots, t_n$ for a fixed smoothing parameter $\lambda > 0$. For each value of λ let us assume that there is an $n \times n$ cubic spline smoother matrix $S_\lambda = (f_\lambda(x_1), f_\lambda(x_2), \cdots, f_\lambda(x_n)^T$ which is positive semi definite.

The estimators satisfying the sum of squares equation (3)

$$\beta_\lambda = (\tilde{X}^T \tilde{X})^{-1} \tilde{X}^T (I - S_\lambda) Y \tag{4}$$

and

$$\hat{Y} = S_\lambda + \tilde{X}(\tilde{X}^T\tilde{X})^{-1}\tilde{X}^T(I - S_\lambda) \tag{5}$$

Select a value for the smoothing parameter λ based on the minimizer of the generalized cross-validation (GCV) criterion

$$GCV(\lambda) = \frac{n(Y - \hat{Y})^T(Y - \hat{Y})}{(n - trH_\lambda)^2} \tag{6}$$

where $tr\mathbf{H}_\lambda$

$$tr\mathbf{H}_\lambda = trS_\lambda + tr(\tilde{X}^T\tilde{X})^{-1}\tilde{X}^T(I - S_\lambda)\tilde{X} \tag{7}$$

The smoothing parameter $\hat{\lambda}$ in the semi–parametric setting can finally be obtained by minimizing the UBR criterion

$$UBR(\lambda) = \frac{1}{n}(\|y - x\|^2 + 2\sigma^2 tr(\mathbf{H}_\lambda)) \tag{8}$$

where $\mathbf{H}_\lambda = S_\lambda + \tilde{X}(\tilde{X}^T\tilde{X})^{-1}\tilde{X}^T(I - S_\lambda)$. $\|\cdot\|$ denotes the Euclidean norm and $\|y - x\| = \|(I - \mathbf{H}_\lambda)y\|$.

2.3 Optimize Smoothing Parameter by GA

At present, the smoothing parameters mainly uses the GCV techniques [14,15]. But the method involves many complex mathematical knowledge and nonlinear optimizes probles, and is difficult for programming. Then it limits its application in actual engineering. The GA is applied to search the optimal smoothing parameter λ, and then we use semi–parameteric regression to estimate the unknown parameters. The fitness function is defined as follows:

$$f(\lambda) = \frac{1}{UBR} = n(\|y - x\|^2 + 2\sigma^2 tr(\mathbf{H}_\lambda)) \tag{9}$$

The optimization procedures are as follows:

1. Select a sample of n and select the break knots of the splines $t_1, t_2, \cdots, t_n$ by stepwise deletion method.
2. Calculate $\mathbf{H}_\lambda$ by the sample information to meet equation (4).
3. The initial group is randomly generated with L individuals, and each individual is made up of binary code string, each represent a chromosome $\lambda \in [0, 100]$.
4. Calculate $tr(\mathbf{H}_\lambda)$ corresponding to k chromosomes, and obtain the values of fitness function and GCV.
5. Retain the individual of the highest adaptation in the group, because it does not participate in crossover and mutation calculation and directly copy it to the next generation. As for other individuals in the group, a roulette selection is adapted. Discarded chromosomes in the last step are replaced with new random chromosomes, so that the initial number n of chromosomes can be maintained. This replacement can reinforce the random search of optimum carried out by mutations. After a small percentage of the gene in the list of chromosomes by means of random mutation.

6. generate groups of a new generation. Repeat the steps (2) − (4), and the group will evolve a generation each time, until the adaptation meets with the requirements or achieve the overall algebra of evolution. The stopping criterion is usually satisfied either the population converged to a unique solution or a maximum number of predetermined generations reached.
7. select 20 individuals of higher mutation from the generation of the last evolution, and twenty better smoothing parameters λ will be got. Then make GCV testing technology, a optimal smoothing parameters λ can get.

3 Experiment Study

In this section, two main financial time series are used to test the proposed semi–parameters regression forecasting model. First of all, we describe the data and evaluation criteria used in this study and then report the experimental results.

3.1 Data Description and Evaluation Criteria

There are more than 100 technical indicators that can be used to attempt to gain insight about market behavior. Some of them model market fluctuations and some focus on when to make buy and sell decisions. The difficulty with technical indicators is deciding which indicators are crucial in determining market movements. It is not wise to use all the available indicators in a model. In this paper, the established prediction model takes into account that the various technical indicators recorded the important act of the market, and in accordance with the conditions in financial market select 12 stock market technical indicators as variables factors which affect the stock market [16]. The description of 8 variables are presented in Table 1.

Table 1. 12 variables and their formulas

Variables indicators	Formulas
Open Prices as Prediction Variable (y)	$x_t, (t = 1, 2, \cdots, n)$
High Price (x_1)	$max(x_t)$
Low Price (x_2)	$min(x_t)$
Stochastic oscillator (SO)(x_3)	$\frac{x_t - x_3(m)}{x_h(m) - x_l(m)}$
Rate of change (ROC)(x_4)	$\frac{x_t}{x_{t-5}}$
Moving Average (MA_5)(x_5)	$\frac{1}{5}\sum_{i=t-5}^{t} x_i$
Moving Variance (MV_5)(x_6)	$\frac{1}{5}\sum_{i=t-5}^{t}(x_i - \bar{x_t})$
Moving Variance Ratio (MVR_5)(x_7)	MV_t^2/MV_{t-5}^2
Disparity5 (D5)(x_8)	x_t/MA_5
Disparity10 (D10)(x_9)	x_t/MA_{10}
Price oscillator (OSCP)(x_{10})	$\frac{MA_5 - MA_{10}}{MA_5}$
Price Oscillator (PO)(x_{11})	$(MA_5 - MA_{10})/MA_5$
Commodity channel index (CCI)(x_{12})	$\frac{M_t - SM_t}{0.0015D_t}$

The data set used for our experiment consists of two time series data: the S&P 500 index series index series. The data used in this study are obtained from Datastream (http://www.datastream. com). The entire data set covers the period from October 27, 2008 to December 31, 2009. This paper established the predictive models based on 240 daily data from October 27, 2008 to October 7, 2009 as the training data sets, and test model based on 60 daily data from October 8, 2009 to December 31, 2009 as the testing data set, which are used to evaluate the good or bad performance of predictions. This paper get 6 composite variables as independent variables by PLS from 12 technical indicators, The open prices y is the dependent variable as prediction.

In order to measure effectiveness of the proposed method, two typical indicators, normalized mean squared error (NMSE) and directional change statistics (D_{stat}) were used in this paper [17]. Given n pairs of the actual values (y_t) and predicted values ($\hat{y}_t$), the NMSE can be defined as

$$NMSE = \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y}_i)^2} \tag{10}$$

where $\bar{y}_i$ is the average of actual value.

Directional change statistics (D_{stat}) can be expressed as

$$D_{stat} = \frac{1}{n}\sum_{t=1}^{N} I_t \cdot 100\% \tag{11}$$

where $I_t = 1$ if $(y_{t+1} - y_t)(\hat{y}_{t+1} - \hat{y}_t) \geq 0$, and $a_t = 0$ otherwise.

3.2 Analysis of the Results

This paper is presented the model of semi-parametric smoothing regression with GA for financial time series forecasting named as SA-GA. In order to investigate the effect of the proposed model, traditional Stepwise Linear Regression (SLR) and RBF neural networks with Gaussian–type activation functions (RBF–NN) are established by the same independent variables. Those are fitted the 240 samples and forecasted the 60 samples by the those models, the comparison results are used to test the effect of predictive models. The standard RBF neural networks with Gaussian–type activation functions in hidden layer were trained for each training set, network was trained using the neural network toolbox provided by Matlab software package.

SR-GA parameters are set as follows: the iteration times are 100; the population is 100; Figure 1 shows the curve of fitness in the training stage. One can see that the best, mean and the worst fitness and convergent speed are tending towards stability with increase of iteration number. Therefore, optimal smoothing parameters λ can get.

Figure 2, 3 and 4 show 300 sample in fitting and forecasting, we can see that the differences between the different models are very significant. The more important factor to measure performance of a method is to check its generalization ability. From Figure, we can see that forecasting results of SR–GA is very best, which has obvious advantages over other two models in the same variables for training samples and testing samples.

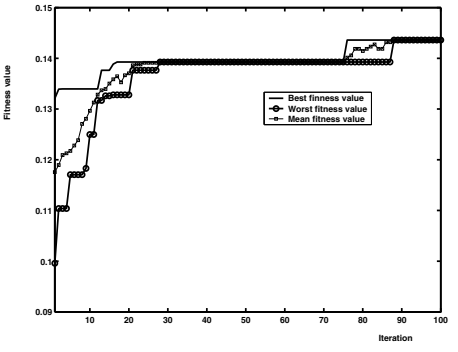


Fig. 1. Fitness Values in Iteration

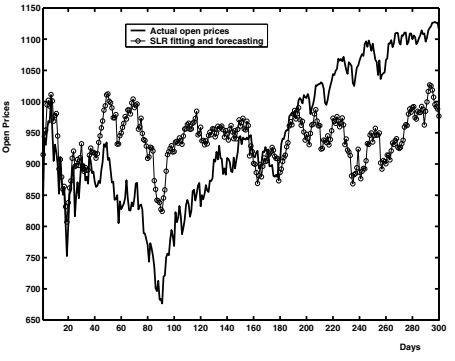


Fig. 2. The results of SLR model

Tables 2 shows the fitting results of 360 training samples for different models. In the table 2, a clear comparison of various methods for the financial time series is given via NMSE and D_{stat}.The results show that SR–GA model better than those of the SLR model and RBF–NN forecasting models for the financial time series in fitting. Tables 4 shows the forecasting results of 60 testing samples for different models about different measure index. Generally speaking, the forecasting results obtained from the tables also indicate that the prediction performance of the proposed SR–GA forecasting model is better than those of the SLR model and RBF–NN forecasting models for the financial time series in forecasting.

Table 2. The NMSE and D_{stat} comparison with different models for 240 training samples

Method	S&P500			
	NMSE	Rank	$D_{stat}(\%)$	Rank
SLR	0.4376	3	53.43%	3
RBF–NN	0.2060	2	98.30%	1
SR–GA	0.1795	1	87.92%	2

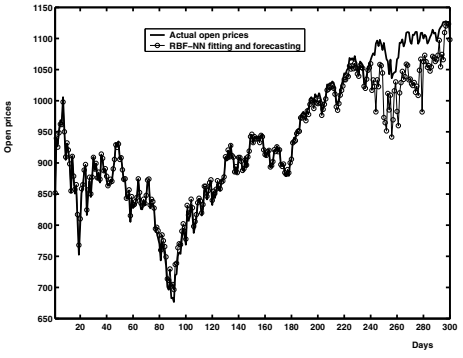


Fig. 3. The results of RBF–NN

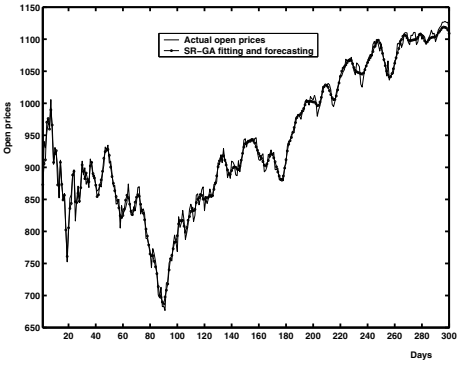


Fig. 4. Comparison of the results of SR–GA

In detail, the NMSE of the our proposed SR–GA model reaches 0.1675 in the testing samples for S & P500 time series, however the NMSE of the SLR model is 5.4790; the NMSE of the RBF–NN model is 7.4021; These results show the NMSE of the SR–GA model is less than other model, which has obvious advantages over other models for S & P500 time series forecasting.

Similarly, for D_{stat} efficiency index, the proposed SR–GA model has higher valuer in the testing samples for S & P500 time series forecasting. From Table 4, the D_{stat} for the SR–GA model reaches only 86.31%, while for the SLR model, the D_{stat} is 56.78%; the D_{stat} of theRBF–NN model is only 32.56%; Those show that the SR–GA model is close to their financial series data for S & P500 time series forecasting. From Tables 3 and 4, the results show the results of RBF–NN model is unstable and over–fitting.

Table 3. The NMSE and D_{stat} comparison with different models for 60 testing samples

Method	S&P500			
	NMSE	Rank	$D_{stat}(\%)$	Rank
SLR	5.4790	2	56.78%	2
RBF–NN	7.4021	3	32.56%	3
SR–GA	0.1675	1	86.31%	1

4 Conclusions

It is very critical to choose smoothness parament λ, which is a tuning parameter which controls the tradeoff between goodness of fitting and complexity of the model. In order to improve the smoothness of curve fitted by the finical time series model of polynomial spline functions, the adaptive semi–parametric

regression with a penalized item is introduced to estimate the unknown parameters. The GA is applied to search the optimal smoothing parameter. Finical time series faces a complex external environment of rapid change. In this study, PLS technology is used to synthesize information, and GA is used to optimization the smooth parament for forecasting model. Examples of calculation shows that the method can significantly improve the system's predictive ability, prediction accuracy, and with a high prediction accuracy of the rising and falling trend of the finical times. Empirical results obtained reveal that the proposed nonlinear combination technique is a very promising approach to financial time series forecasting. And the empirical results also confirm that. Therefore, it is very significant to establish a kind of model that can improve not only fitting accuracy but also fitting smoothness for theoretical research and practical applications.

Acknowledgment. The authors would like to express their sincere thanks to the editor and anonymous reviewer's comments and suggestions for the improvement of this paper. This paper is supported by the National Natural Science Foundation of China under Grant No. 11161029.

References

1. Francis, E.H., Chao, L.J.: Modified support vector machine in finaancial time series forecasting. Neurocomputing 48, 847–861 (2002)
2. Zhang, G., Patuwo, B.E., Hu, M.Y.: Forecasting with artificial neural networks: the state of the art. International Journal Forecasting 14, 35–62 (1998)
3. Oh, K.J., Kim, K.: Analyzing stock market tick data using piecewise nonlinear model. Expert Systems with Applications 22, 249–252 (2002)
4. Huang, W., Lai, K.K.: Forecasting foreign exchange rates with artificial neural networks: a review. International Journal of Information Technology & Decision Making 3, 145–165 (2004)
5. Majhi, R., Panda, G., Sahoo, G.: Efficient prediction of exchange rate with low complexity artificial neural network models. Expert Systems with Application 36, 181–189 (2009)
6. Nott, D.: Semiparametric estimation of mean and variance functions for non-Gaussian data. Computational Statistics 21, 603–620 (2006)
7. Ruppert, D., Wand, M.P., Carroll, R.J.: Semiparametric Regression. Cambridge University Press, New York (2003)
8. Wu, J.: A Semi-parametric Regression Ensemble Model for Rainfall Forecasting Based on RBF Neural Network. In: Wang, F.L., Deng, H., Gao, Y., Lei, J. (eds.) AICI 2010, Part II. LNCS(LNAI), vol. 6320, pp. 284–292. Springer, Heidelberg (2010)
9. Tutz, G.: Generalized semiparametrically structured mixed models. Computational Statistics and Data Analysis 46, 777–800 (2004)
10. Wu, J., Liu, M.Z., Jin, L.: A Hybrid Support Vector Regression Approach for Rainfall Forecasting Using Particle Swarm Optimization and Projection Pursuit Technology. International Journal of Computational Intelligence and Applications 9(3), 87–104 (2010)

11. Luo, F., Wu, J., Yan, K.: A novel nonlinear combination model based on support vector machine for stock market prediction. In: Proceedings of the 8th World Congress on Intelligent Control and Automation, Jinan, China, pp. 5048–5053 (2010)
12. Wu, J.: A novel artificial neural network ensemble model based on K–nn nonparametric estimation of regression function and its application for rainfall forecasting. In: Yu, L., Lai, K.K., Mishra, S.K. (eds.) Proeedings of the 2nd Internatioal Joint Conference on Computational Sciences and Optimization, vol. 2, pp. 44–48. IEEE Computer Society Press (2009)
13. Wu, J., Chen, E.: A Novel Nonparametric Regression Ensemble for Rainfall Forecasting Using Particle Swarm Optimization Technique Coupled with Artificial Neural Network. In: Yu, W., He, H., Zhang, N. (eds.) ISNN 2009, Part II. LNCS, vol. 5553, pp. 49–58. Springer, Heidelberg (2009)
14. Herrmann, E.: Variance estimation and bandwidth selection for kernel regression. In: Schimek, M.G. (ed.) Smoothing and Regression. Approaches, Computation and Application, pp. 71–107. John Wiley, New York (2000)
15. Kohn, R., Schimek, M.G., Smith, M.: Spline and kernel smoothing for dependent data. In: Schimek, M.G. (ed.) Smoothing and Regression. Approaches, Computation and Application, pp. 135–158. John Wiley, New York (2000)
16. Kim, K.J.: Financial time series forecasting using support vector machines. Neurocomputing 55, 307–319 (2003)
17. Christoffersen, P.F., Diebold, F.X.: Financial asset returns, direction-of-change forecasting, and volatility dynamics. Management Science 5, 1273–1287 (2006)

Classification of Respiratory Abnormalities Using Adaptive Neuro Fuzzy Inference System

Mythili Asaithambi[1], Sujatha C. Manoharan[2], and Srinivasan Subramanian[1]

[1] Department of Instrumentation Engg., MIT Campus, AnnaUniversity, Chennai, India
[2] Department of Electronics and Communication Engineering,
CEG Campus, Anna University, Chennai, India
mythiliasaithambi@gmail.com

Abstract. Spirometric evaluation of pulmonary function plays a critical role in the diagnosis, differentiation and management of respiratory disorders. In spirometry, there is a requirement that a large database is to be analyzed by the physician for effective investigation. Hence, there is a need for automated evaluation of spirometric parameters to diagnose respiratory abnormalities in order to ease the work of the physician. In this work, a neuro fuzzy based Adaptive Neuro Fuzzy Inference System (ANFIS), Multiple ANFIS and Complex valued ANFIS models are employed in classifying the spirometric data. Four different membership functions which include triangular, trapezoidal, Gaussian and Gbell are employed in these classification models. Results show that all the models are capable of classifying respiratory abnormalities. Also, it is observed that CANFIS model with Gaussian membership function performs better than other models and achieved higher accuracy. This study seems to be clinically relevant as this could be useful for mass screening of respiratory diseases.

Keywords: Flow-volume spirometry, forced expiratory maneuver, Adaptive Neuro Fuzzy Inference System, Multiple Adaptive Neuro Fuzzy Inference System, Complex-valued Adaptive Neuro Fuzzy Inference System.

1 Introduction

Spirometry is a valuable tool for the primary care clinician when making respiratory diagnoses, assessing progress and predicting prognosis. Various international guideline updates for respiratory disorders have stated that spirometry is mandatory in order to confirm the diagnosis of chronic obstructive pulmonary disease and other related diseases [1]. Spirometric pulmonary function test can detect the presence of airflow obstruction, lung volume reduction and is also useful in distinguishing respiratory diseases from cardiac diseases. Thus, Spirometry is established as an essential diagnostic tool and can be performed in most physicians' offices, with the patient sitting comfortably in front of the spirometer [9].

Spirometry is a measure of dynamic changes in lung volumes and capacities as a function of time during forced expiration and inspiration. The spirogram, a flow-volume pattern curve, is characterized by various flow and volume parameters during

J.-S. Pan, S.-M. Chen, N.T. Nguyen (Eds.): ACIIDS 2012, Part III, LNAI 7198, pp. 65–73, 2012.

various time intervals. Respiratory diseases and their severity are interpreted based on these patterns and the parameters obtained from them [2,3]. Some of the important parameters obtained from the maneuver are Forced Expiratory Volume of air in one second (FEV_1), Forced Vital Capacity (FVC) and Peak Expiratory Flow (PEF). FEV_1 is the volume of air exhaled in one second, FVC is the total amount of air exhaled and PEF is the maximum flow achieved during the expiration. Automated diagnosis methods that evaluate respiratory flow volume patterns, and classify spirometric data, for appropriate interpretation are very much essential [6-8].

Artificial Neural Networks (ANN) have already been used in the classification of respiratory data [3]. ANN employed for classification problems do not guarantee high accuracy, besides being computationally heavy [4]. The necessity for a large training set to achieve high accuracy is another drawback of ANN. On the other hand, fuzzy logic technique which promises better accuracy depends heavily on expert knowledge, that may not always be available [5]. Even though it requires less convergence time, it depends on trial and error method in selecting the fuzzy membership functions. These problems are overcome by the hybrid neuro fuzzy model which removes the stringent requirements as it includes the benefits of both ANN and the fuzzy logic systems [4]. Successful implementations of ANFIS have been reported for classification, data analysis [10-12] of many biomedical engineering applications such as detection of breast cancer [16], analysis of EEG and ECG signals [17,18].

Multiple ANFIS, which is an extension of the ANFIS network, has been employed for human facial expression recognition [19] and multiple objective decision making problem [20]. Complex valued ANFIS model is based on Sugeno fuzzy system. Each fuzzy rule of this model is realized by a neural network. CANFIS has been widely used for classification problems [21, 22]. In this work, the classification of spirometric data has been demonstrated using ANFIS, MANFIS and CANFIS models and their performance are compared.

2 Methodology

The spirometer recordings are carried out on adult volunteers (N = 250) for the present study. The age, gender and race of the subject are recorded and height, weight are being measured before recording. The portable Spirolab II spirometer with a gold standard volumetric transducer is used for the acquisition of the data.

ANFIS model used in this work is based on Sugeno fuzzy model [11]. This multilayer neural network based fuzzy system generates fuzzy rules and parameters of Membership Function (MF) from a given input data set. Each rule in the ANFIS model is formed as:

$$\text{IF } x_1 \text{ is } A_{1,j} \quad \text{AND} \quad x_2 \quad \text{is } A_{2,j} \quad \text{AND ... AND } x_n \quad \text{is } A_{n,j} \tag{1}$$

$$\text{THEN } y = c_0 \quad + c_1 x_1 \quad + c_2 x_2 \quad + \ldots + c_n x_n \tag{2}$$

where $A_{i,j}$ is the j^{th} linguistic term of the i^{th} input variable x_i, and n is the number of inputs. y is the output variable and c_i are consequent parameters to be determined in the training process.

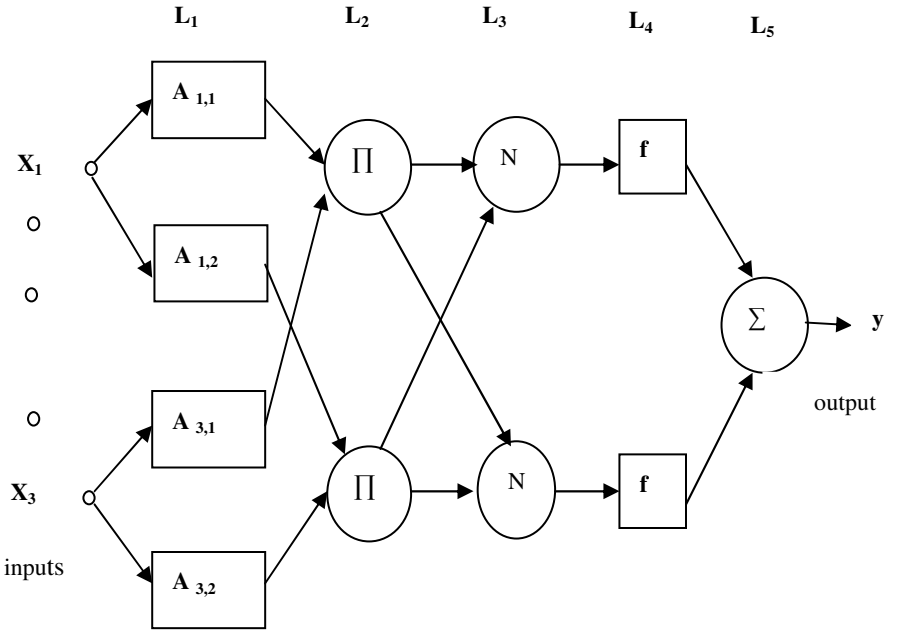


Fig. 1. Architecture of ANFIS

The architecture of the ANFIS model consists of five layers with their associated nodes and is shown in Fig. 1.

Layer 1 represents the membership functions. In this layer, there are n × k nodes, where n is the number of the input variables and k is the number of membership functions. The MF maps each input element to a membership grade value between 0 and 1.

Every node in this layer 2 is a fixed node labelled , whose output is the product of all the incoming signals. Each node output represents the firing strength of a rule and it is given by:

$$w_i = \prod \mu_{Ai}(x_j), i = 1,2;\ j = 1,2,...,n \tag{3}$$

where w_i represents multiplication of the input values from the previous layer, with respect to firing strength of the i^{th} rule.

Layer 3 is the normalization layer where the rule strength is normalised as:

$$\overline{w}_i = \frac{w_i}{\sum w_i} \tag{4}$$

where w_i is the firing strength of the i^{th} rule.

Layer 4 is an adaptive layer. Every node in this layer is a linear function and the coefficients of the function are adapted through a combination of least squares approximation and back propagation.

$$\overline{w}_i f_i = \overline{w}_i (c_0 + c_1 x_1 + c_2 x_2 ..... + c_n x_n) \tag{5}$$

Layer 5 is the output layer which computes overall output as the summation of all incoming signals. The output is computed as:

$$\sum_i \overline{w}_i f_i = \frac{\sum w_i f_i}{\sum w_i} \tag{6}$$

where $\overline{w_i f_i}$ is the output of the node i in layer 4. The overall output is linear, even though the premise parameters are nonlinear. The values of input and output nodes of ANFIS represent the training values and the predicted values respectively.

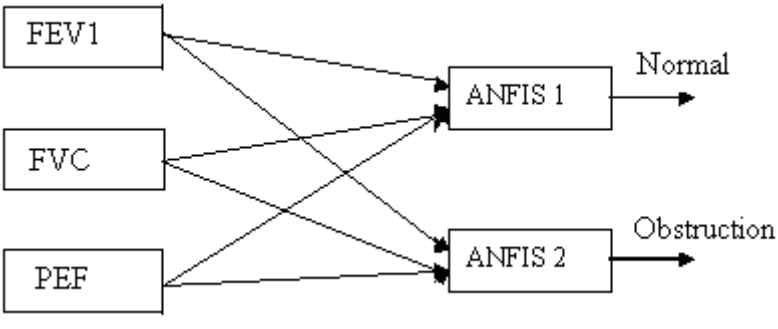


Fig. 2. Architecture of the MANFIS

MANFIS, which is an extension of ANFIS has two ANFIS placed in parallel to differentiate the normal from abnormal values. This architecture serves as a basis for constructing a set of fuzzy if-then rules with appropriate membership functions to generate the stipulated input-output pairs. The output of each rule is a linear combination of input variables plus a constant term, and the final output is the weighted average of each rule's output. In the forward pass of the learning algorithm, the consequent parameters are identified by the least square estimates. In the backward pass the premise parameters are updated by gradient descent.

In MANFIS, each ANFIS has an independent set of fuzzy rules, which makes it difficult to realize the possible correlations between outputs. The parameters of the membership functions are not shared by the ANFIS models. Also, the number of adjustable parameters drastically increases as number of output increases. In Complex valued ANFIS model the weight normalization and the membership function values are shared to express the correlations. This interaction between the membership functions during training is considered as an added advantage. [22].

In this work, the spirometric parameters FVC, FEV_1 and PEF are given as inputs to the various models. Membership functions which include, Bell-shaped function, Gaussian function, triangular-shaped function and trapezoidal shaped functions are used in the classification process. The functions are defined as:

$$\text{Bell-shaped function } \mu_{Ai} = \frac{1}{1+\left|\frac{x-c}{a}\right|^{2b}} \tag{7}$$

$$\text{Gaussian function } \mu_{Ai} = \exp\left\{-\left[\frac{x-c}{\sigma}\right]^2\right\} \tag{8}$$

where c and σ are centres and widths of the function and are referred to as the antecedent parameters for the membership function

$$\text{Triangular function } \mu_{Ai} = \max\left(\min\left(\frac{x-a}{b-a}, \frac{c-x}{c-b}\right), 0\right) \tag{9}$$

where a, b, c are antecedent parameters, a, c locate the feet of the triangle and b locates the peak.

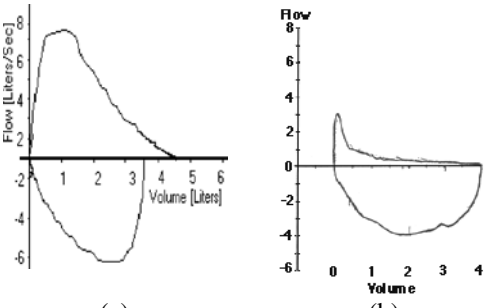
$$\text{Trapezoid function } \mu_{Ai} = \max\left(\min\left(\frac{x-a}{b-a}, 1, \frac{d-x}{d-c}\right), 0\right) \tag{10}$$

where a, c are antecedent parameters locate the feet of the trapezoidal and b, d locates the top. These parameters are adaptive and they determine the value of the i^{th} membership function for each variable to the fuzzy set A_i. As the values of the parameters change, the value of the function also varies accordingly, thereby indicating the degree to which the input variable x satisfies the membership function. The optimal number of membership functions are chosen based on the values of classification accuracy. Receiver Operating Characteristic (ROC) analysis has been used to plot sensitivity with respect to specificity to measure the accuracy of the classifier [14].

3 Results and Discussion

The spirometric pulmonary function test is investigated using flow-volume spirometer. In this study, flow-volume data are recorded for N=250 and the parameters such as FEV_1, FVC and PEF are obtained from them. Fig 3(a) shows a typical response of a spirometer that depicts variation of airflow with lung volume for a normal subject. It is seen that this normal flow-volume curve has rapidly rising flow at the beginning of expiration and then it declines steadily in a linear fashion. The

highest flow rate, obtained during the first part of expiration, is effort dependent and is approximately equal to one third of vital capacity.



(a) (b)

Fig. 3. (a) and (b). Typical flow-volume curves of normal and abnormal subjects

Fig 3(b) shows a typical flow-volume curve of subject with obstruction abnormality. There is a rapid peak expiratory flow but the curve descends more quickly than normal and takes on a concave shape. The classifiers are trained with a set of 180 data and their performance is estimated by computing specificity, sensitivity and accuracy for test data of 70 subjects.

Table 1. Performance estimators for different Membership functions for ANFIS model

MF type	Performance estimator		
	Sensitivity	Specificity	Accuracy
Triangular	78.78	93.3	88.17
Trapezoidal	69.44	90.74	82.22
Gaussian	77.42	89.83	85.56
Gbell	81.25	93.1	88.89

Table 2. Performance estimators for different Membership functions for MANFIS model

MF type	Performance estimator		
	Sensitivity	Specificity	Accuracy
Triangular	92.86	93.55	93.33
Trapezoidal	93.33	98.1	96.67
Gaussian	79.31	88.52	85.56
Gbell	83.33	84.85	84.44

Table 3. Performance estimators for different Membership functions for CANFIS model

MF type	Performance estimator		
	Sensitivity	Specificity	Accuracy
Triangular	93.15	95.4	91.25
Trapezoidal	95	97.2	90.82
Gaussian	92.5	98	97.55
Gbell	85.3	86	82.5

The performance measures of ANFIS, MANFIS and CANFIS classification models with different types of membership functions are presented in Tables. 1, 2 and 3. Table. 1, the values of sensitivity of ANFIS model are found to be high for all the membership functions. This shows that this model is able to identify abnormal subjects well. It also observed that high values of classification accuracy, specificity and sensitivity are obtained for ANFIS model with Gaussian membership function.

From Table.2, it is found that the values of sensitivity for MANFIS model with various membership functions are greater than that of ANFIS model. It is also observed that this model with trapezoidal membership function achieved higher values of accuracy, sensitivity and specificity than MANFIS model utilizing other membership functions.

It is found that CANFIS model with Gaussian membership function performs better in classification as shown in Table.3. It is also further observed that CANFIS model with Gaussian membership function obtained higher values of classification accuracy, sensitivity and specificity than MANFIS and ANFIS model.

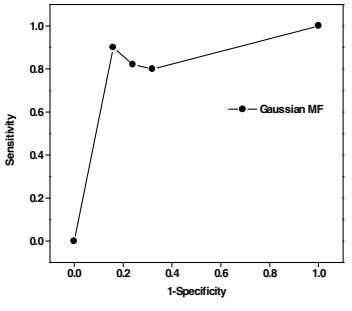


Fig. 4. ROC analyses for Gaussian membership function of CANFIS

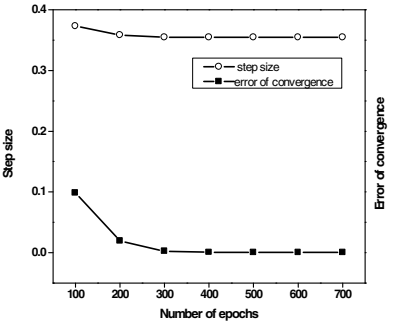


Fig. 5. Adaptation of parameter steps and error of convergence of CANFIS

The ROC curves for CANFIS model with varied number of membership functions (three, five and seven) is shown in Fig 4. Results demonstrate that CANFIS model with three number of Gaussian membership functions achieved better classification accuracy.

The steps of parameter adaptation and the network error convergence curve of CANFIS are shown in Figure 5. The step size had an initial value of 0.1 and converged to 0.0002 after the 200 epochs. Similarly, the network error decreases with increase in number of epochs and converged to constant value at higher epochs.

4 Conclusion

Spirometry is the most frequently performed pulmonary function test and is an essential tool for the diagnosis of respiratory diseases. The clinical utility of spirometer depends on the accuracy, performance of the subject and on the measured

and predicted values [5]. It is also reported that a large database is to be analyzed by the physician to investigate the pulmonary function abnormalities. Hence there is a need to provide automated diagnostic support to the physician using hybrid intelligent systems.

In this work, ANFIS, MANFIS and CANFIS models with various membership functions are employed for classification of spirometric data. It has already been shown that automated analysis of spirometric pulmonary function data is carried over using neural networks [15]. ANFIS model achieves higher classification accuracy when compared to the previously employed neural network model. This could be due to the reason that ANFIS combines both the learning capabilities of neural networks and reasoning abilities of fuzzy inference system. Also, it is found that CANFIS obtains higher classification accuracy of 97.5% compared to ANFIS and MANFIS model. The network convergence error was low with three number of membership functions. Results demonstrate that the proposed model can be used to enhance the diagnostic relevance of pulmonary function test. This method can be used for automated mass screening and further enhanced in severity classification of respiratory abnormalities.

Acknowledgement. The authors gratefully acknowledge Department of Science and Technology (DST), Government of India, New Delhi for financial support under PURSE scheme.

References

1. Robin, C., Vicky, T., Gareth, W.: Impact of an Educational Intervention on the Quality of Spirometry Performance in a General Practice: an Audit. Primary Care Respiratory Journal 20(2), 210–213 (2011)
2. Thomas, L.P.: Benefits of and barriers to the widespread use of spirometry. Current Opinion in Pulmonary Medicine 11, 115–120 (2005)
3. Sujatha, C.M., Mahesh, V., Ramakrishnan, S.: Comparison of two ANN methods for classification of spirometer data. Measurement Science Review 8(3), 53–57 (2008)
4. Jude, H.D., Kezi, S.V.C., Anitha, J.: Application of Neuro-Fuzzy Model for MR Brain Tumor Image Classification. Biomedical Soft Computing and Human Sciences 16(1), 95–102 (2009)
5. Uncu, U.: Evaluation of Pulmonary Function Tests by Using Fuzzy Logic Theory. Journal of Medical Systems 34(3), 241–250 (2010)
6. Derom, E., Van Weel, C., Listro, G., Buffels, J., Schermer, T., Lammers, E., Wouter, E., Decramer, M.: Primary care spirometry. European Respiratory Journal 31(1), 197–203 (2008)
7. 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)
8. 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)

9. David, P., Daryl, F., Jen, C., Alan, K., Frank, C.: Earlier diagnosis and earlier treatment of COPD in primary care. Primary Care Respiratory Journal 20(1), 15–22 (2011)
10. Abdurrahim, A., Serkan, K., Niyazi, K., Osman, N.U., Nilgun, A.: Diagnosis of Renal Failure Disease Using Adaptive Neuro-Fuzzy Inference System. Journal of Medical Systems 34(6), 1003–1009 (2010)
11. Loukas, Y.L.: Adaptive neuro-fuzzy inference system: an instant and architecture-free predictor for improved QSAR studies. J. Med. Chem. 44, 2772–2783 (2001)
12. Cheng, H.W., Baw, J.L., Lawrence, S.H.W.: The Association Forecasting of 13 Variants Within Seven Asthma Susceptibility Genes on 3 Serum IgE Groups in Taiwanese Population by Integrating of Adaptive Neuro-fuzzy Inference System (ANFIS) and Classification Analysis Methods. Journal of Medical Systems (2010), doi:10.1007/s10916-010-9457-4
13. Sujatha, C.M., Ramakrishnan, S.: Prediction of Forced Expiratory Volume in Pulmonary function test using Radial basis Neural Networks and k-means Clustering. Journal of Medical Systems 33(5), 347–351 (2009)
14. Centor, R.M.: Signal Detectability: The use of ROC curves and their analysis. Medical Decision Making 11(2), 102–106 (1991)
15. Sujatha, C.M., Ramakrishnan, S.: Prediction of forced expiratory volume in normal and restrictive respiratory functions using spirometry and self-organizing map. Journal of Medical Engineering & Technology 33(7), 538–543 (2009)
16. Elif, D.U.: Adaptive Neuro-Fuzzy Inference Systems for Automatic Detection of Breast Cancer. Journal of Medical System 33(5), 353–358 (2009)
17. Guler, N.: Application of adaptive neuro-fuzzy inference system for detection of electrocardiographic changes in patients with partial epilepsy using feature extraction. Expert Systems with Applications 27(3), 323–330 (2004)
18. Inan, G., Elif, D.U.: Adaptive neuro-fuzzy inference system for classification of EEG signals using wavelet coefficients. Journal of Neuroscience Methods, NSM-3945 (2005)
19. Gomathi, V., Ramar, K., Santhiyaku, A.J.: Human Facial Expression Recognition using MANFIS Model. Proceedings of World Academy of Science Engineering and Technology 38, 338–342 (2009)
20. Chi-Bin, C., Cheng, C.J., Lee, E.S.: Neuro-Fuzzy and Genetic Algorithm in Multiple Response Optimization. PERGAMON Computers and Mathematics with Applications 44, 1503–1514 (2002)
21. Eiji, M., Kenichi, N.: Modular neural network-type CANFIS neuro-fuzzy modeling for multi-illumination color device characterization. In: Joint 9th IFSA World Congress and 20th NAFIPS International Conference, vol. 4, pp. 2090–2095 (2001), 2001)
22. Alireza, M.-A., Mohammad-R, A.-T.: Complex-Valued Adaptive Neuro Fuzzy Inference System-CANFIS. In: Proceedings of the 2004 World Automation Congress and Fifth International Symposium on Soft Computing for Industry, Spain (2004)

An Ant Colony Optimization and Bayesian Network Structure Application for the Asymmetric Traveling Salesman Problem

Nai-Hua Chen

Department of Information Management, Chienkuo Technology Univeristy,
No.1, Chiehshou North Road, Changhua City 500, Taiwan
nhc@cc.ctu.edu.tw

Abstract. The asymmetric traveling salesman problem (ATSP) is an NP-hard problem. The Bayesian network structure which describes conditional independence among subsets of variables is useful in reasoning uncertainty. The ATSP is formed as the Bayesian network structure and solved by the ant colony optimization (ACO) in this study. The proposed algorithm is tested in different sample size. The exam case is finding customer preference's city sequence. Results show the proposed algorithm has a higher joint probability than random selected case. More applications such as the sequential decision, the variable ordering or the route planning can also implement.

Keywords: Ant colony optimization, Bayesian network, Asymmetric Traveling Salesman Problem.

1 Introduction

The development of information technology grows large data sets. A powerful data analysis tool is essential to extract useful knowledge from large data sets for supporting decision making. Bayesian network structure which describes conditional independence among subsets of variables is useful in reasoning uncertainty. The basic concept of the Bayesian network combines prior knowledge and observed data to probabilistic prediction. An optimal decision making can be provided from the analysis result of the Bayesian network.

The traveling salesman problem (TSP) is finding the shortest path of visiting each city once. The TSP is a well know NP-hard optimization problem. Many researchers use the heuristic algorithm to approach this problem. In recently, researches develop heuristic algorithms inspired by mother natural to solve optimal problems. Some popular heuristic algorithms include simulated annealing (AN) [5], genetic algorithms (GA) [6], particle swarm optimization (PSO) [8] and ant colony optimization (ACO) [7] are used in optimization problem.

The ACO algorithm was to mimic the ant behavior in finding the shortest path to food. The concept of the ACO is close to TSP problem and was first design in solving the TSP problem.

J.-S. Pan, S.-M. Chen, N.T. Nguyen (Eds.): ACIIDS 2012, Part III, LNAI 7198, pp. 74–78, 2012.

This research adopts the ACO and the Bayesian network structure in finding the ATSP algorithm has a better performance. The rest of the paper is organized as follows. Section 2 introduces some related topics the research background. The approach problem is described in section 3. The algorithm is tested by different sample size and compared with a radon chosen sequence. Experimental results are presented in section 4. Section 5 is ended with conclusions.

2 Research Background

2.1 The Asymmetric Traveling Salesman Problem (ATSP)

The ATSP is a well-known NP-hard optimization problem. The ATSP is defined on a directed graph G= (V, A), where $V=\{V_i, i=1\ldots,n\}$ is the set of vertices and $A=\{(V_i,V_j),\forall V_i,V_j\in V\}$ is denoted as the set of direct arcs that connect all vertices. Let c_{ij} be the distance from the vertex V_i and the vertex V_j. The value of x_{ij} is set as 1 if the arc (V_i,V_j) is connected, otherwise, the value of x_{ij} is set as 0. The mathematical formulation of the traditional ATSP can be written as follows (Öncan et al., 2009).

$$\min \sum_{i=1}^{n}\sum_{j=1}^{n} c_{ij}x_{ij} \tag{1}$$

$$s.t.\sum_{i=1}^{n} x_{ij} = 1,\quad j=1,\cdots,n, \tag{2}$$

$$\sum_{j=1}^{n} x_{ij} = 1,\quad i=1,\cdots,n, \tag{3}$$

$$0 \le x_{ij} \le 1,\quad i,j=1,\cdots,n, \tag{4}$$

$$x_{ij} \in \{0,\ 1\},\quad i,j=1,\cdots,n, \tag{5}$$

$$\{(i,j): x_{ij}=1,\ i,j=2,\cdots,n\} \text{ does not contain subtours.} \tag{6}$$

The formulations of (2) and (3) indicate that each arc has one incoming arc and one outgoing arc. The tour start from a fixed node, say 1, is written as the constraint (6). The c_{ij} is associated with the traversal cost of the arc (V_i, V_j) and is represented as the length generally. The graph is asymmetric therefore the cost c_{ij} and the cost c_{ji} are different. In this research, the cost defined based on the Baysian network structure.

2.2 The Bayesian Network

The Bayesian Network is acyclic probabilistic graph model. Each notes represents a random variable and arrows signify the existence of direct caused influenced from parents notes [2]. The general form of the joint probability distribution can be expressed as:

$$P(V_1,\ldots,V_n)=\prod_i P(V_i\,|Pa(V_i)) \quad ,$$

where A={V_1,…,V_n} represents the system variable, Pa(V_i) is the parent set of the vertex V_i . Many researches use the Bayesian network applications in the decision making of the social science. The graph of one direct traveling salesman problem is shown in Figure 1 and the according joint probability is presented as $P(V_1,V_2,V_3,V_4,V_5) = P(V_5 \mid V_4)P(V_4 \mid V_3)P(V_3 \mid V_2)P(V_2 \mid V_1)$.

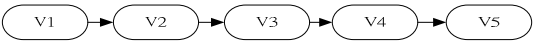


Fig. 1. A direct traveling salesman problem graph

2.3 Ant Colony Optimization (ACO)

The ACO is a meta heuristic algorithm and was proposed by Dorigo in 1990 [1]. The algorithm is inspired from real ant colonies in finding food. As ants find food sources and they deposit pheromone on the trail. Other ants happened to choose the trail with strong amount of pheromone deposited by previous ants. The communication behavior direct ants find a shortest path between the nest and the food. The ACO algorithm was first used to solve the traveling salesperson problem (TSP). In the ACO algorithm, ants initial choose a city randomly. The next city to visit depends on the proportional rule. The according probability of ant k travel from city i to city j is expressed as

$$p_{ij}^{k} = \frac{(\tau_{ij})^{\alpha}(\eta_{ij})^{\beta}}{\sum_{N_i^k}(\tau_{ij})^{\alpha}(\eta_{ij})^{\beta}}, \text{ if } j \in N_i^k$$

where τ_{ij} denotes the level of pheromone deposited in the route from city i to city j. The η_{ij} is the heuristic information. In the TSP problem, $\eta_{ij}=1/d_{ij}$, d_{ij} is the distance between city i and city j. The$\alpha$and the$\beta$are two parameters which determine the relative influence of the pheromone trail and the heuristic information. The N_i^k is the feasible neighborhood that ant k when being at city i. After evaporation, all ants deposit pheromone one the trip, the pheromone is update as

$$\tau_{ij} \leftarrow (1-\rho)\tau_{ij} + \sum_{k=1}^{m}\Delta\tau_{ij}^{k}$$

where $\Delta\tau_{ij}^{k}$ is the amount of pheromone of ant k deposits when traveling from city i to city j. The $\Delta\tau_{ij}^{k}$ is defined as

$$\Delta\tau_{ij}^{k} = \begin{cases} 1/L^{k}, & \text{if the trip from city i to city j is complete by ant k} \\ 0, & \text{otherwise} \end{cases}$$

where L^k is the trip length from city i to city j. Finally, the artificial ants build a shortest route to visit each city sequentially.

3 The Approach Problem

Finding the shortest path for the ATSP has been applied in real-world problems, such as scheduling, route planning, and some cost reduction etc. Recently, the shortest path or the lowest cost is not the only main issues to gain industries profits. The customer preference is another solution for increasing industries profits. This research tries to find the strong linkage of the path for ATSP. The object of the problem is defined as

$$\max P(V_1,...,V_n) = \Pi_{i=1}^{n-1} P(V_{i+1} \mid V_i),$$

where $P(V_1,\ldots, V_n)$ is the joint probability of the path.

4 Computational Experience and Performance

The sequential suggestion list for the eco-tour spotlights is applied in this study. A survey analysis is first used to evaluate customers' preference for the eco-tour spotlights. Respondents are asked if they prefer to visit 74 selected eco-tour spotlights. Results are used for build a Bayesian network structure. In the ACO algorithm for the TSP, the two factors guide the move of ants which are trails and attractiveness [4]. The trails means amount of pheromone of travel, which is referred as τ_{ij} in the ACO algorithm. The τ_{ij} is update by a conditional probability, $P_k(V_j| V_i)$, when the ant k move from city i to city j as follows.

$$\tau_{ij} \leftarrow \tau_{ij} \times \Pi_k P_k(V_j \mid V_i).$$

The attractiveness is associated with the quality of the node which is referred as η_{ij} in this paper. In this study, the η_{ij} is replaced as the joint probability of city i and city j, $\eta_{ij} = P(V_i \cap V_j)$. The algorithm of adopting ACO for predicting recommend sequences is as shown as follows.

```
program ACOATSP
Loop
Initial the pheromone of each city as τ_i0 =P(Vi)
All artificial ants are positioned in cities randomly
  For step= 1: number of city
    For k=1: number of ants
      Choose the next city to move base on the
probability
```

$$p_{ij}^{k} = \frac{(\tau_{ij})^{\alpha}(\eta_{ij})^{\beta}}{\sum_{N_i^k}(\tau_{ij})^{\alpha}(\eta_{ij})^{\beta}}$$

```
    End For
  End For
Update the pheromone trail as τ_ij ← τ_ij × Π_k P_k(V_j | V_i)
Until End Condition
end.
```

The coefficient of the α, β in the ACO algorithm is set as 2 and 1 respectively. The modified ACO algorithm is compared with randomly selected sequence. Table 1 shows the joint probability of the proposed algorithm and the random selected model with different sample size of 20, 40, 60 and 80. Results shows that the proposed algorithm has better performance.

Table 1. The joint probability of proposed and random selected sequence

Sample size	Modified ACO	Random Selected
20	4.5911e-006	9.7739e-008
40	1.5537e-012	2.5395e-015
60	1.0355e-018	1.1238e-022
80	8.9827e-025	4.1184e-030

5 Conclusion

The algorithm for solving the ATSP is developed in this study. We use the ACO and Bayesian network structure to find the maximum joint probability. A random selected sequence is also compared with our proposed algorithm. Results show that our proposed algorithm can get a better solution. More applications such as the variable ordering, the sequence decision or the route planning can also implement based on this algorithm.

References

1. Dorigo, M., Stützle, T.: Ant Colony Optimization. The MIT Press, Cambridge (2004)
2. Jensen, F.V.: Introduction to Bayesian Networks. Springer, Berlin (1996)
3. Pearl, J.: Probabilistic Reasoning in Intelligent System: Networks of Plausible Inference. Morgan Kaufmann, San Mateo (1988)
4. Shukla, A., Tiwari, R., Kala, R.: Real Life Applications of Soft Computing. CRC press, Taylor & Francis, Boca Raton, FL (2010)
5. Kirkpatrick, S., Gelatt Jr., C.D., Vecchi, M.P.: Optimization by Simulated Annealing. Sci. 220(4598), 671–680 (1983)
6. Holland, J.H.: Adaptation in Natural and Artificial Systems. U. of Michigan Press (1975)
7. Dorigo, M.: Optimization, Learning and Natural Algorithms (Ph.D Thesis). Politecnico di Milano, Italie (1992)
8. Kennedy, J., Eberhart, R.: Particle Swarm Optimization. In: Proc. of IEEE Intl. Conf. on Neural Networks IV, pp. 1942–1948 (1995)
9. Öncan, T., Altinel, I.K., Laporte, G.: A comparative analysis of several asymmetric traveling salesman problem formulations. Comp. & Oper. Res. 36, 637–654 (2009)

Expert Pruning Based on Genetic Algorithm in Regression Problems

S.A. Jafari[1,5], S.Mashohor[2], Abd. R. Ramli[3], and M. Hamiruce Marhaban[4]

[1,2,3] Department of Computer and Communication Systems Engineering, Faculty of Engineering, University Putra Malaysia
[4] Department of Electrical and Electronics Engineering, Faculty of Engineering, University Putra Malaysia
[5] Petroleum Training Center of Mahmoudabad, Mazandaran, Iran
sajkenari@gmail.com,
{syamsiah,arr,hamiruce}@eng.upm.edu.my

Abstract. Committee machines are a set of experts that their outputs are combined to improve the performance of the whole system which tend to grow into unnecessarily large size in most of the time. This can lead to extra memory usage, computational costs, and occasional decreases in effectiveness. Expert pruning is an intermediate technique to search for a good subset of all members before combining them. In this paper we studied an expert pruning method based on genetic algorithm to prune regression members. The proposed algorithm searches to find a best subset of experts by creating a logical weight for each member and chooses which member that the related weight is equal to one. The final weights for selected experts are calculated by genetic algorithm method. The results showed that MSE and R-square for the pruned CM are 0.148 and 0.9032 respectively that are reasonable rather than all experts separately.

Keywords: expert pruning, committee machine, learning algorithms, genetic algorithm.

1 Introduction

The ensembles made by existing methods are sometimes needlessly large. The disadvantages of ensembles are; using the extra memory, the computational overhead, and the occasional decreases in effectiveness. There are also some individuals with low predictive performances that create negative effect on overall performance of an ensemble. Pruning committee members while preserving a high diversity among the remaining individuals is an efficient technique for increasing the predictive performance. In the other word, the most advantage of expert pruning is efficiency and predictive performance. In fact the expert pruning problem is similar as an optimization problem, which the objective is finding the best subset of individuals from the original Committee Machine (CM). As we know, in a set with N member there

J.-S. Pan, S.-M. Chen, N.T. Nguyen (Eds.): ACIIDS 2012, Part III, LNAI 7198, pp. 79–88, 2012.

is $(2^N - 1)$ subset, so in a CM with a moderate size the exhaustive search becomes intractable. In [1], authors proposed a clustering method for ensemble pruning. In their method, experts are divided to some sets according to their outputs similarity and then one single network from each cluster is selected. This method does not guarantee that the selected experts improve the generality prediction of the ensemble. In [2], the authors presented a pruning approach based on a Genetic Algorithm (GA) named GASEN (Genetic Algorithm based Selective Ensemble). At the first step, GASEN trains a number of neural networks and then assigned random or equal weights to those networks. At the second step, the framework employed GA to change the weights so that the optimum weights of the neural networks in constituting an ensemble with lowest error can be tackled. Finally the networks whose weight is bigger than a preset threshold λ could be selected to join the ensemble and the models of the ensemble that did not exhibit the predefined threshold are dropped. After selecting the ensemble members the new weights for all candidates can be obtained by normalizing the oldest weight or reapplying the GA to sub ensemble. In [3], the authors extended the technique of Stacked Regression to prune the members of the ensemble using an equation including both accuracy and diversity. The diversity is based on measuring the positive correlation between the errors of the ensemble members and the accuracy of each individual is calculated relative to the accuracy of the most accurate ensemble member. A drawback of this method is predefinition of a weighting GA parameter to balance accuracy and diversity which has to define by user. More details of GA can be found in [4] and [5].

In this paper we studied an expert pruning or ensemble pruning method that can be used to prune regression committee members. The proposed algorithm aims to search for a best subset of experts by making a logical weight for each expert and finally chooses the best experts which the related weight value is one. The rest of this paper is structured as follows. Section 2 will presents related work on ensemble pruning in regression problems. In Section 3 we will introduce our methodology which is based on GA. Data set preparation, experts design with different training algorithms, expert pruning, combine the obtained subset, results and discussion are presented in Section 4. In Section 5, conclusion will be summarized.

2 Related Works

In this section we introduce some important proposed methods in ensemble pruning that are applied in regression problems. The explanation will explore three type of methods which are; Directed Hill Climbing (DHC), Semi Definite Programming (SDP) and ordered aggregation.

2.1 Directed Hill Climbing Method (DHC)

Hill climbing search greedily moves from current state to the next state, which is in its neighborhood. Let $H = \{h_t; t = 1,2,...,T\}$ be an original ensemble with T members and $S \subseteq T$ is a sub ensemble. At first, the method selects an empty (or full) initial sub

ensemble S and continues with searching in the space of different ensembles by iteratively expanding (or contracting) the set S by a single member h_t. The search is guided by an evaluation measure that is the main component of the hill climbing algorithm which proposed in different methods by some authors. The experimental results obtained by different hill climbing methods report good predictive performance results [6-8]. Figure 1; illustrate the search space for an ensemble with four members. One of the important parameters in designing DHC method is the direction of search that is in two types which are; forward selection and backward elimination.

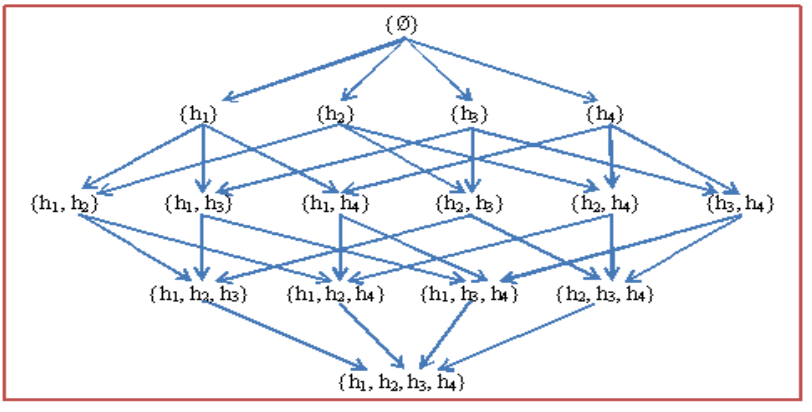


Fig. 1. An example of forward search in DHC

In forward selection, firstly, the sub ensemble S is initialized to the empty set and then the algorithm continues by iteratively adding the individual $h_t \in H \setminus S$ to set S so that optimizes the predefined evaluation function. In backward elimination, the sub ensemble S is equal with *H* at the start and then the algorithm continues by iteratively remove the individual $h_t \in S$ to optimize the evaluation function [6]. The second significant parameter in designing DHC method is the type of evaluation measure that are based on performance and diversity. In performance basis, the objective function is defined based on increasing the performance of the produced ensemble which is created by adding or removing a model to or from the current ensemble [7, 9]. In [10, 11], the authors have used diversity as an evaluation measure. Finding a suitable calculating method for diversity during the search of sub ensemble plays very significant role and needs much more attention. The third significant parameter is the evaluation of data sets. The evaluation function scores the candidate sub ensemble according to its diversity or accuracy. These procedures need a set of data for performance which will be called the as pruning set but it is clear that training set or separate validation can also be used as pruning set. In [12], the authors used a k-fold cross validation such that the remaining fold of the training set are used to create an ensemble. The same fold is used as the pruning set for models and sub ensembles of the ensemble. Finally, the evaluations are averaged across all folds. Because the evaluation of models is based on unseen data that were not used for their training, therefore this method

is less prone to over fitting. The last important parameter is amount of pruning that is the size of the final sub ensemble which can be determine in two ways. One way is using the fixed number or fixed percentage for sub ensemble size which can be defined by user before starting the test. Another way is dynamic population size which is based on predictive performance of the sub ensemble with different size. In this approach, the performance of the whole sub ensemble scores during the search from initial population to final in forward or backward selection will be analyzed. Finally sub ensemble with best performance will be selected as final ensemble which successfully improves the efficiency.

2.2 Semi Definite Programming (SDP)

In [13], the authors proposed a pruning method based on Semi Definite Programming (SDP) for classification tasks, and in [14], the author used this method for regression problem. They formulated an ensemble pruning problem as a quadratic programming to obtain sub ensemble classifiers with optimal accuracy and diversity trade off. At first, they defined an error matrix P, which records the misclassification of each classifier on the training set as follows:

$$P_{ij} = \begin{cases} 0, & \text{if } jth \text{ classifier on data point } i \text{ is correct} \\ 1, & \text{otherwise} \end{cases} \tag{1}$$

Let $G = P^t P$, which the diagonal arrays G_{ii} are the whole errors made by classifier i and the term G_{ij} is the whole errors made by classifier i and j. The goal is to find a sub matrix of G, such that the sums of the elements in row or column are minimum. After normalization, the matrix $\tilde{G}$ will be:

$$\tilde{G} = \begin{cases} G_{ii} / N & \text{if } i = j \\ \frac{1}{2}\left(\frac{G_{ij}}{G_{ii}} + \frac{G_{ij}}{G_{jj}} \right) & \text{if } i \neq j \end{cases} \tag{2}$$

Where N is number of training data and whole array of matrix in $\tilde{G}$ are belong to (0, 1). The diagonal array (G_{ii}) are the error rate of classifier i and the off-diagonal array (G_{ij}) are the overlap of errors between classifier i and j. Note that G_{ij} / G_{ii} is the conditional probability that classifier j misclassifies a point, given that classifier i does. Taking the average of G_{ij} / G_{ii} and G_{ij} / G_{jj} as the off-diagonal elements of $\tilde{G}$ makes the matrix symmetric. Naturally, sum of diagonal array in matrix $\tilde{G}$ measures the overall strength and sum of off-diagonal array measures the diversity of the ensemble. Then a combination of diagonal and off-diagonal array in $\tilde{G}$ should be a good approximation of the ensemble error and the ensemble is good if the whole array is small values. The equation is a quadratic integer programming formula which can be

used for the subset selection problem with a fixed-size subset of classifiers (k), and the objective function is the sum of the corresponding elements in the $\tilde{G}$ matrix that should be minimized.

$$\begin{aligned} &\min \ w^t \tilde{G} w \\ &\text{s.t.} \begin{cases} \sum_i w_i = k \\ w_i \in \{0,1\} \end{cases} \end{aligned} \tag{3}$$

The weight w_i represents whether *i-th* classifier is included in the sub ensemble (when $w_i = 1$) or no ($w_i = 0$). If $w_i = 1$, it means that the corresponding diagonal and off-diagonal array will be counted in the objective function and vice versa. The equation is a standard 0-1 optimization problem, which is NP-hard in general and therefore they formulate it as a "max cut problem" with size k. In graph theory the max cut problem is a method that partitions the vertices of an edge weighted graph into two sets with same size k, so that the total weight of edges crossing the partition is maximized.

2.3 Ordered Aggregation

In practice, committee members are trained sequentially and then collected as they are created from the different methods. In this method of creating an ensemble, by increasing the ensemble member, the individual error shows a monotonic decreasing. Ordered aggregation is a method in ensemble pruning which reorders the members of an original ensemble and then selects a sub ensemble. In the past two decades a number of researchers have sought to determine sub ensemble by ordered aggregation methods [7, 11, 14]. In [15], authors used ordered aggregation technique to prune a regression bagging ensemble. From the initial pool of M predictors generated by bagging, ordered aggregation builds a sequence of nested ensembles, in which the sub ensemble of size u contains the sub ensemble of size $(u - 1)$. Such as forward selection in hill climbing method, the algorithm starts with an empty set of experts and grows by adding a new expert in each iteration such that the new set reduces the training error of the extended sub ensemble. In this method, the algorithm creates a variance-covariance matrix C that has to be minimized. For example in iteration u, the new *k-th* expert is selected to add to the sub ensemble. The matrix $C(u)$ will be obtained by adding all covariance between last $(u-1)$ experts and the new expert k with the variance of this expert. In other words, the matrix $C(u)$ is created by the matrix $C(u-1)$ and all variance-covariance between new expert and the oldest experts. The objective function is the sum all array in matrix $C(u)$ which has to be less than the sum of array in matrix $C(u-1)$. In this situation, the expert k is suitable to be added to the set of $(u-1)$ experts to create new sub ensemble. This algorithm repeats until the whole of experts are tested. In [16], the authors proposed an AdaBoost algorithm using reweighting to compute the weighted training error and ordering them with considering the aggregation of the members generated by the Bagging method. Their

objective was modifying the aggregation order in a bagging ensemble. Finally they used the AdaBoost weighting method to compute the training error and the expert with the lowest weighted error is selected from a pool of experts generated by bagging. This process can be applied to any other parallel ensemble building methods.

3 Methodology

In this paper we proposed a pruning method based on GA that has been used for regression problems. The objective in any regression problem is to learn a predictor (expert) of the dependent variable $y \in R$ (in one dimensional output) as a function of the independent variables $X = (x_1, x_2, ..., x_n) \in R^n$ (attributes) using the training data $D = \{(x_i, y_i), i = 1,2,...,n\}$ that is drawn from a probability distribution $P(D)$. Assume $\hat{f}_i(x)$ is the prediction given by the ith expert on sample data *D*. The final predicted output of the CM with N members is a combination of the individual output with the weights w_i that are showed as below:

$$F_{CM}(x) = \sum_{i=1}^{N} w_i \hat{f}_i(x); \qquad 0 \le w_i \le 1 \quad \& \quad \sum_{i=1}^{N} w_i = 1 \tag{4}$$

The error of the CM is

$$E = \int (F(x) - f(x))^2 P(x) dx \tag{5}$$

Where $f(x)$ is target function to approximate and $P(x)$ is the probability density distribution in attribute space. To prune the regression CM we try to select a sub-CM with M members that minimize the error function which has been introduced in Equation (5). For selection of the sub-CM we use mean square error based on the training error and expect this estimation to be similar to calculation that is done over the true distribution of the data. Then we assume that minimizing the error given by training data leads to the minimization of the generalization error. Actually in real regression problem, the training error minimization usually leads to over fitting. Indeed, the experiments carried out show that the size of the sub-CM based on training error tend to be smaller than the optimal sub-CMs based on test data [15]. With assumption of a subset of original CM with lower generalization error, the process of finding this subset is complex and needs generating the whole $(2^N - 1)$ non empty sub-CM. In literature, finding an optimal subset of members that minimizes the error estimated on data set is defined as the NP-hard problem and is not generally feasible in practice. Our proposed method is a practical approach to expert pruning based on GA to find out whose experts that should be excluded from the CM. The main idea behind this proposed method is heuristics, i.e. assuming each expert can be assigned a weight that would characterize the fitness function, and then the experts whose weight is equal to one could be selected to join the sub-CM. Suppose that the weight of the *i-th* member of committee is w_i, which satisfies in below equation where *K* is the cardinality of sub-CM which is user defined.

$$w_i \in \{0,1\}, \qquad \sum_{i=1}^{N} w_i = K \tag{6}$$

Then we get a weight vector $W = (w_1, w_2, ..., w_N)$ that w_i is equal to zero or one for each $i = 1, 2, ..., N$. This weight vector should minimize the generalization error of the CM which is Mean Square Error (MSE) as a fitness function in GA. If $w_i = 0$, that means *i-th* expert is not belong to our final subset and if $w_i = 1$, which means the *i-th* expert is belong to our final pruned CM. Then after minimizing the objective function we have K weight, which are equal to one and $(N - K)$ weight, which are equal to zero. Each weight that is equal to one means the related expert is belong to our final pruned CM. Then we have K members for final sub-CM and they can be combined with different method and get a new weight for them that satisfies the Equation (4).

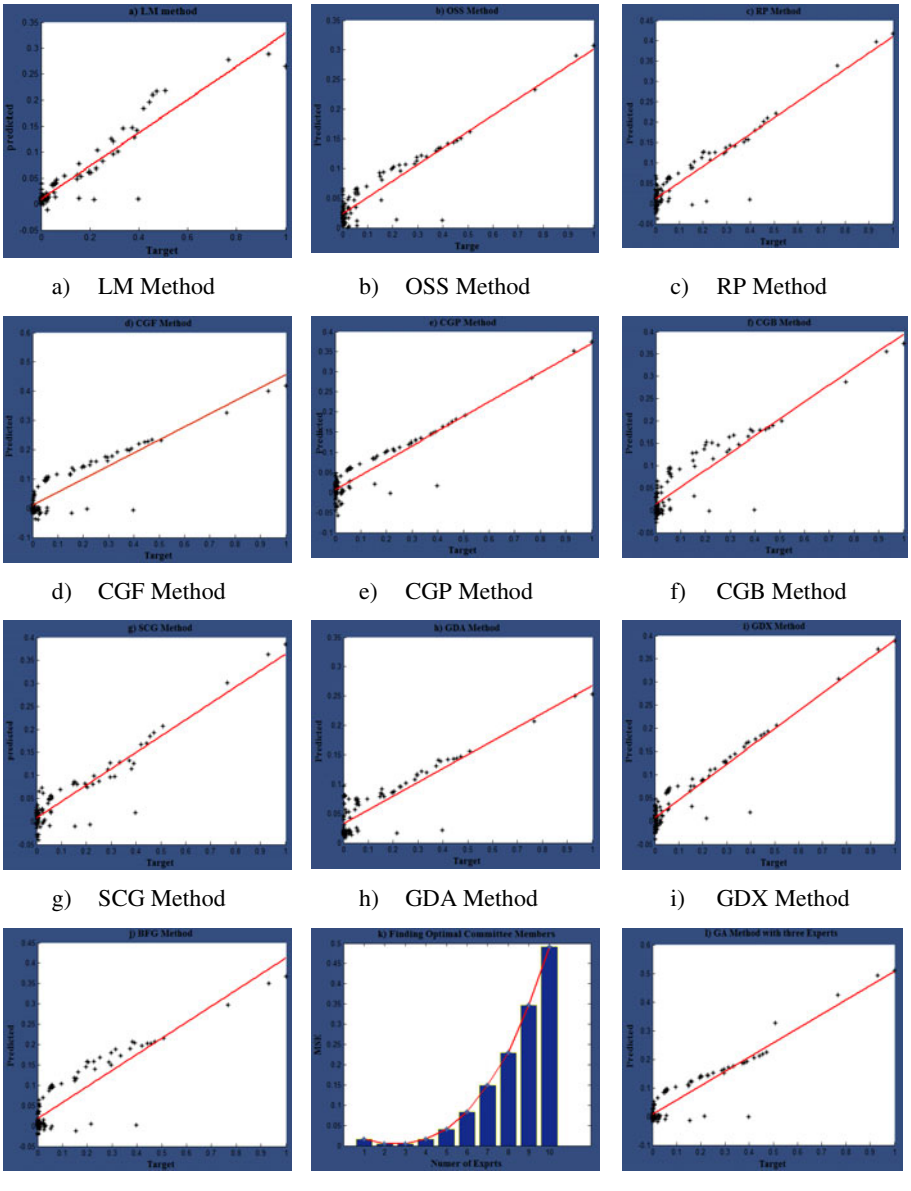
4 Simulation Results and Discussion

The simulation is performed under Matlab on a personal computer with Intel core 2 duo 2.7 GHZ processor and 4 GB of memory. We carried out experiments over a set of oil data from National Iranian Oil Company (NIOC) that were well log data and core permeability data. Sonic transit time (DT), Neutron log (NPHI), Density log (RHOB), Gamma Ray (GA), and True Formation Resistivity (Rt) are the input data and rock permeability (K) is a target data. Ten neural networks with different training algorithms have been used as committee members that were: Variable learning rate back propagation (GDA, GDX), Resilient back propagation (RP), Conjugate gradient (CGF, CGP, CGB, SCG), Quasi-Newton (BFG, OSS), Levenberg-Marquardt (LM). Figure 2 (a-j) illustrated the cross plot between target and predicted value by different training algorithms for testing data. After that the proposed method for expert pruning are implemented on the initial CM. The performance of the GA method on all individuals showed that three experts can decrease the error compared with other subset of the initial CM. Figure 2(k) illustrates the MSE versus the number of experts. Therefore by applying our proposed method, experts with no. 3, 4 and 10 are selected as the best subset. On average, it takes about 30 minutes in each iteration. After finding the optimal subset, the final weights for these three experts are calculated by GA method which are 0.25, 0.6 and 0.15 respectively. Table 1, illustrates MSE and correlation coefficient for all ten experts and the final weight for three selected experts based on GA method for making the final sub CM.

Table 1. MSE and R^2 for ten experts and the final GA calculated weight for selected experts

Method no.	(1) LM	(2) OSS	(3) RP	(4) CGF	(5) CGP	(6) CGB	(7) SCG	(8) GDA	(9) GDX	(10) BFG
MSE	0.0023	0.024	0.018	0.016	0.02	0.019	0.021	0.026	0.019	0.018
R^2	0.887	0.868	0.882	0.82	0.867	0.82	0.876	0.779	0.897	0.804
Weight	---	---	0.25	0.6	---	---	---	---	---	0.15

According to the results, the pruned CM has produced the minimum error and reasonable correlation coefficient rather than all experts that are 0.148 for MSE and 0.9032 for R-square. Figure (2-l) shows the scatter plot of target and predicted final output with GA method as a combination method for sub CM.



a) LM Method b) OSS Method c) RP Method

d) CGF Method e) CGP Method f) CGB Method

g) SCG Method h) GDA Method i) GDX Method

j) BFG Method k) Different subset l) Output of sub CM

Fig. 2. The performance of different sub CM (k); The crossplot between target value and predicted value for ten learning algorithms (a)-(j) and for combined sub CM with GA (l)

5 Conclusion

In this paper we presented an expert pruning method based on GA. Ten neural networks with different training algorithms are created as experts for initial CM. After that the proposed pruning method is applied to find the optimal sub CM. As mentioned in section 2, the size of the final subset can be defined by user before starting the test. In this paper we tried to find optimal subset size by adding one by one member to initial empty subset. In this approach, the subset with best performance is selected as final sub CM which successfully improved the MSE as evaluation measure. As illustrated in figure 2(k) the subset with three experts is the optimal subset for our investigation that are RP, CGF and BFG learning algorithms respectively. After finding the optimal subset, the final weights for these three experts are calculated by GA method which are 0.25, 0.6 and 0.15 respectively (refer table 1). Mean square error and correlation coefficient obtained by applying GA in final sub CM are 0.148 and 0.9032 respectively. Therefore in comparing with MSE and R-square obtained for all members, these sub CM results appears more reasonable. Figure (2-l) shows the scatter plot of measured and predicted final output with GA method as a combination method for sub CM. The advantage of the proposed method related to DHC, SDP and OA that mentioned in section 2, is the similarity computation in our method that is very significant for optimal subset selection.

Acknowledgements. The authors wish to express their appreciation to the National Iranian Oil Company (NIOC) for sponsorship, data preparation and their collaboration. Also, we would like to thank engineer Mr. .Hamid Mosalmannejad from Iranian Offshore Oil Company (IOOC) for his invaluable experiences shared during this research.

References

1. Bakker, B., Heskes, T.: Clustering ensembles of neural network models. Neural Netw. 16(2), 261–269 (2003)
2. Zhou, Z.-H., Wu, J., Tang, W.: Ensembling neural networks: Many could be better than all. Artificial Intelligence 137(1-2), 239–263 (2002)
3. Rooney, N., Patterson, D., Nugent, C.: Reduced Ensemble Size Stacking. In: 6th IEEE International Conference on Tools with Artificial Intelligence, pp. 266–271 (2004)
4. Patel, R., Shrawankar, U.N., Raghuwanshi, M.M.: Genetic Algorithm with Histogram Construction Technique. In: Second International Conference on Emerging Trends in Engineering & Technology, pp. 615–618 (2009)
5. Huang, H.-C., Chen, Y.-H.: Genetic fingerprinting for copyright protection of multicast media. Soft Computing 13(4), 383–391 (2008)
6. Banfield, R.E., et al.: Ensemble diversity measures and their application to thinning. Information Fusion 6(1), 49–62 (2005)
7. Caruana, R., et al.: Ensemble selection from libraries of models. In: Proceedings of the Twenty-First International Conference on Machine Learning (2004)
8. Partalas, I., Tsoumakas, G., Vlahavas, I.: An ensemble uncertainty aware measure for directed hill climbing ensemble pruning. Machine Learning 81(3), 257–282 (2010)

9. Fan, W., et al.: Pruning and dynamic scheduling of cost-sensitive ensembles. In: Eighteenth National Conference on Artificial Intelligence, pp. 146–151 (2002)
10. Brown, G., Wyatt, J.L., Peter: Managing Diversity in Regression Ensembles. Machine Learning Research 6, 1621–1650 (2005)
11. Martinez-Munoz, G., Suarez, A.: Pruning in ordered bagging ensembles. In: Proceedings of the 23rd International Conference on Machine Learning, pp. 609–616 (2006)
12. Caruana, R., Munson, A., Niculescu-Mizil, A.: Getting the most out of ensemble selection. In: International Conference on Data Mining, pp. 828–833 (2006)
13. Zhang, Y., Burer, S., Street, W.N.: Ensemble Pruning Via Semi-definite Programming. Jornal of Machin. Learnning Research 7, 1315–1338 (2006)
14. Martinez-Muoz, G., Hernandez-Lobato, D., Suarez, A.: An Analysis of Ensemble Pruning Techniques Based on Ordered Aggregation. IEEE Transactions on Pattern Analysis and Machine Intelligence 31(2), 245–259 (2009)
15. Hernandez-Lobato, D., Martinez-Munoz, G., Suarez, A.: Pruning in Ordered Regression Bagging Ensembles. In: Proceedings of the IEEE World Congress on Computational Intelligence, pp. 1266–1273 (2006b)
16. Martínez-Muñoz, G., Suárez, A.: Using boosting to prune bagging ensembles. Pattern Recognition Letters 28(1), 156–165 (2007)

A Hybrid CS/PSO Algorithm for Global Optimization

Amirhossein Ghodrati[1] and Shahriar Lotfi[2]

[1] Computer Engineering Department, College of Nabi Akram, Tabriz, Iran
[2] Computer Science Department, University of Tabriz, Tabriz, Iran
a.h.ghodrati@gmail.com, shahriar_lotfi@tabrizu.ac.ir

Abstract. This paper presents the hybrid approach of two nature inspired metaheuristic algorithms; Cuckoo Search (CS) and Particle Swarm Optimization (PSO) for solving optimization problems. Cuckoo birds lay their own eggs to other host birds. If the host birds discover the alien birds, they will leave the nest or throw the egg away. Cuckoo birds migrate to the environments that reduce the chance of their eggs to be discovered by the host birds. In standard CS, cuckoo birds experience new places by the Lévy Flight. In the proposed hybrid algorithm, cuckoo birds are aware of each other positions and make use of swarm intelligence in PSO in order to reach to better solutions. Experimental results are examined with some standard benchmark functions and the results show a promising performance of this algorithm.

Keywords: Cuckoo Search, Global optimization, PSO, Metaheuristic and Hybrid evolutionary algorithm.

1 Introduction

Finding optimal solutions for many problems is very difficult to deal with. The complexity of such problems makes it impossible to look for every possible solution or combination [1]. However, because of their complexity the use of approximation algorithms in order to find approximate solutions is getting more popular in the past few years [2]. Among these algorithms, modern metaheuristics are becoming popular, which leads to a new branch of optimization, named metaheuristic optimization. Most of these algorithms are nature inspired [3], some of which have been proposed for optimization problems, for example, Genetic Algorithm (GA) [4], Harmony Search (HS) [5], Ant Colony Optimization (ACO) [6], Imperialist Competitive Algorithm (ICA) [7] and Artificial Bee Colony [8].

Yang and Deb formulated a new metaheuristic algorithm, called Cuckoo Search in 2009 [9]. This algorithm is inspired by life of cuckoo bird in combination with Lévy Flight behavior of some birds and fruit flies. The studies show the CS algorithm is very promising and could outperform some known algorithms, such as PSO and GA.

PSO was formulated by Kennedy and Eberhart in 1995 [10]. It is an evolutionary computation technique which is inspired by social behavior of swarms. This algorithm is the simulation of the social behavior of birds, like the choreography of a bird flock. Each individual in the population is a particle and gets a random value in

J.-S. Pan, S.-M. Chen, N.T. Nguyen (Eds.): ACIIDS 2012, Part III, LNAI 7198, pp. 89–98, 2012.

the initialization. Each particle despite the position vector contains the best personal experience and a velocity vector. The particle's velocity and its best personal experience and the global best position all together determines a particle next movement. PSO has variety of usages such as [11] and [12].

CS and PSO are metaheuristic algorithms and are inspired by birds. In this paper cuckoo birds communicate in order to inform each other from the suitable place for laying egg. This is achieved by adding the swarm intelligence which is used in PSO.

Some related works is presented in section 2. Standard CS is discussed in section 3. Section 4 provides an overview of PSO. Section 5 gives a hybrid approach of CS with PSO and Section 6 presents the detailed experimental results to compare the performance of the proposed algorithm with other algorithms.

2 Related Works

Layeb introduced a new hybridization between quantum inspired and cuckoo search for knapsack problem [13]. This hybrid algorithm achieves better balance between exploration and exploitation and experimental results shows convincing results.

A modified CS was proposed by Tuba, Subotic and Stanarevic in 2011 by biasing the step size in the original algorithm [14]. This is achieved by determining the step size from the sorted rather than only permuted fitness matrix.

Another modification of the standard Cuckoo Search was made by Walton, Hassan, Morgan and Brown with the aim to speed up convergence [15]. The modification involves the additional step of information exchange between the top eggs. It was shown that Modified Cuckoo Search (MCS) outperforms the standard cuckoo search and other algorithms, in terms of convergence rates, when applied to a series of standard optimization benchmark objective functions.

Valian, Mohanna and Tavakoli applied cuckoo search to train neural networks with improved performance [16] and was employed for benchmark classification problems and the results shows the effectiveness of the introduce algorithm.

3 Cuckoo Search

This algorithm is inspired by the special lifestyle of a bird called 'cuckoo'. Cuckoo birds never build their own nests and instead lay their eggs in the nest of other host birds. If host birds discover the eggs are not their own, they will either throw these alien eggs away or simply abandon its nest and build a new nest elsewhere. On the other hand cuckoo birds carefully mimic the color and pattern of the eggs of host birds. In general, the cuckoo eggs hatch slightly earlier than their host eggs. Once the first cuckoo chick is hatched, the first instinct action it will take is to evict the host eggs by blindly propelling the eggs out of the nest, which increases the cuckoo chick's share of food provided by its host bird. Studies also show that a cuckoo chick can also mimic the call of host chicks to gain access to more feeding opportunity [9].

In nature, animals' path for searching food is in a random way which effectively is a random walk because the next move is based on the current location and the

transition probability to the next location. Various studies have shown that fruit flies or Drosophila melanogaster; explore their landscape using a series of straight flight path punctuated by a sudden 90° turn, leading to a Lévy-flight-style pattern. Such behavior has been applied to optimization and optimal search, and preliminary results show its promising capability [9], [17]. Figure 1 shows an example of the Lévy flights path [18].

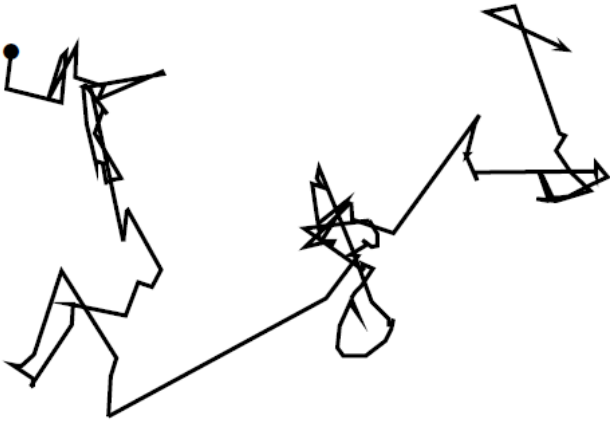


Fig. 1. Example of Lévy flights path [18]

Cuckoo search is introduced in three idealized rules: 1) Each cuckoo lays one egg at a time and dumps it in a randomly chosen nest. 2) The best nest with high quality of eggs (solutions) will carry over to the next generation. 3) The number of available host nests is fixed, and a host can discover an alien egg with a probability $P_a \in [0,1]$.

For maximization problem the fitness of a solution can be proportional to the value of its objective function. Other forms of fitness can be defined in a similar way to the fitness function in other evolutionary algorithm. A simple representation where one egg in a nest represents a solution and a cuckoo egg represents a new solution is used here. The aim is to use the new and potentially better solutions (cuckoos) to replace worse solutions that are in the nests. When generating new solutions $x^{(t+1)}$ for, say cuckoo i, a Lévy flight is performed using the following equation:

$$x_i^{(t+1)} = x_i^t + \alpha \oplus Lévy(\lambda) . \tag{1}$$

Where $\alpha > 0$ is the step size which should be related to the scales of the problem of interest. The product $\oplus$ means entry-wise multiplication. The above equation is essentially the stochastic equation for random walk. In general, a random walk is a Markov chain whose next status/location only depends on the current location (the first term in the above equation) and the transition probability (the second term).

The Lévy flight essentially provides a random walk while the random step length is drawn from a Lévy distribution which has an infinite variance with an infinite mean.

$$Lévy \sim u = t^{-\lambda}, \qquad -1 < \lambda < 3 . \tag{2}$$

Studies show that Lévy flights can maximize the efficiency of resource searches in uncertain environments [18]. Here the consecutive jumps/steps of a cuckoo essentially form a random walk process which obeys a power-law step-length distribution with a heavy tail.

4 Particle Swarm Optimization

Swarm Intelligence (SI) is an innovative distributed intelligent paradigm for solving optimization problems that originally took its inspiration from the biological examples by swarming, flocking and herding phenomena in vertebrates [19].

PSO is an evolutionary algorithm which is inspired by the feeding birds or fish and proposed by Kennedy and Eberhart in 1955. This algorithm like other evolutionary algorithms starts with random initial solutions and begins the process of finding the global optimum. In this algorithm, we call each solution a *particle.* Each particle moves around in the search space with a velocity. The best position explored for a particle so far is recorded and is called *pbest*. Moreover, each particle knows the best *pbest* among all the particles which is called *gbest.* By considering *pbest, gbest* and the velocity of each particle the update rule for their position is as the following equations:

$$V_{t+1} = W_t * V_t + C_1 * rand(\ \ ) * (pbest - x_t) + C_2 * rand(\ \ ) * (gbest - x_t) . \tag{3}$$

$$x_{t+1} = x_t + V_{t+1} . \tag{4}$$

Where W is inertia weight which shows the effect of previous velocity vector (V_t) on the new vector, C_1 and C_2 are acceleration constants and $rand(\ \ )$ is a random function in the range [0, 1] and x_t is current position of the particle.

5 Hybrid CS/PSO Algorithm

In this section, we explore the details of the proposed hybrid algorithm. As mentioned in section 3, the nature of cuckoo birds is that they do not raise their own eggs and never build their own nests, instead they lay their eggs in the nest of other host birds. If the alien egg is discovered by the host bird, it will either throw these alien eggs away or simply abandon its nest and build a new nest elsewhere. Thus cuckoo birds always are looking for a better place in order to decrease the chance of their eggs to be discovered. In the proposed hybrid algorithm, the ability of communication for cuckoo birds has been added. The goal of this communication is to inform each other from their position and help each other to immigrate to a better place. Each cuckoo bird will record the best personal experience as *pbest* during its own life. In addition,

the best *pbest* among all the birds is called *gbest*. The cuckoo birds' communication is established through the *pbest*, *gbest* and they update their position using these parameters and also the velocity of each one. The update rule for cuckoo i's position is as the following:

$$V_{t+1}{}^{i} = W_t{}^{i} * V_t{}^{i} + C_1 * rand(\ ) * \left(pbest - x_t^i\right) + C_2 * rand(\ ) * \left(gbest - x_t^i\right). \quad (5)$$

$$x_{t+1}^i = x_t^i + V_{t+1}{}^{i}\ . \quad (6)$$

Where W is inertia weight which shows the effect of previous velocity vector ($V_t{}^{i}$) on the new vector, C_1 and C_2 are acceleration constants and $rand(\ )$ is a random function in the range [0, 1] and x_t^i is current position of the cuckoo. The pseudo-code of the CS/PSO is presented as bellow:

```
begin
  Objective function f(x), x = (x_1, ..., x_d)^T;
  Initial a population of n host nests x_i (i = 1,2, ..., n);
  While (t<MaxGeneration) or (stop criterion);
    Get a cuckoo (say i) randomly by Lévy flights
    and record pbest;
    Evaluate its quality/fitness F_i;
    Choose a nest among n (say j) randomly;
    if (F_i > F_j),
      Replace j by the new solution;
    end
    Move cuckoo birds using equation 5 and 6;
    Abandon a fraction (P_a) of worse nests
      [and build new ones at new locations via Lévy
      flights];
    Keep the best solutions (or nests with quality
    solutions);
    Rank the solutions and find the current best;
  end while
  Post process results and visualization;
end
```

6 Evaluation and Experimental Results

In this section, we show experimental results which evaluate our hybrid CS/PSO proposed algorithm. In the field of evolutionary computation, it is common to compare different algorithms using benchmark functions, especially when the test involves function optimization [20]. Thus, we applied the proposed algorithm for

some classical benchmark functions which is shown in table 1. Initial range, formulation, characteristics and the dimensions of these problems are listed in table 1. If a function has more than one local optimum, this function is called multimodal. Multimodal functions are used to test the ability of algorithms escaping from local minima. A *p*-variable separable function can be expressed as the sum of *p* functions of one variable. Non-separable functions have interrelation among their variables. Therefore, non-separable functions are more difficult than the separable functions [20]. In table 1, characteristics of each function are given in the third column. In this column, *M* means that the function is multimodal, while *U* means that the function is unimodal. If the function is separable, abbreviation *S* is used to indicate this specification. Letter *N* refers that the function is non-separable.

Table 1. The Benchmark functions

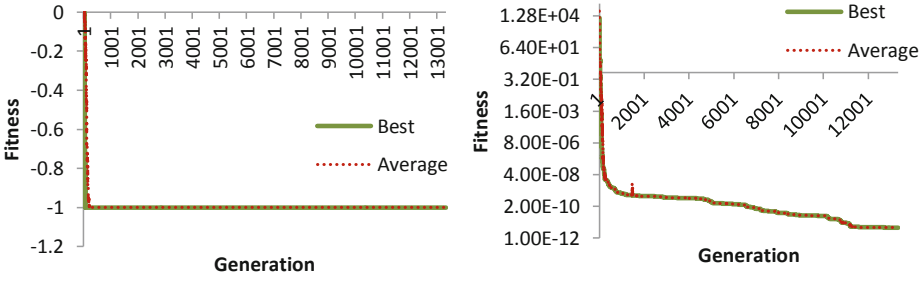
Function	Formulation	Type	DIM	Range
F1 (Ackley)	$-20e^{-0.2\sqrt{\frac{1}{D}\sum_{i=1}^{D}x_i^2}} - e^{\frac{1}{D}\sum_{i=1}^{D}\cos(2\pi x_i)} + 20 + e$	MN	30	[-32,32]
F2 (Dixon-Price)	$(x_1-1)^2 + \sum_{i=2}^{n} i(2x_i^2 - x_{i-1})^2$	UN	30	[-10,10]
F3 (Easom)	$-\cos(x_1)\cos(x_2)\exp(-(x_1-\pi)^2 - (x_2-\pi)^2)$	UN	2	[-100,100]
F4 (Griewank)	$1 + \sum_{i=1}^{D}\left(\frac{x_i^2}{4000}\right) - \prod_{i=1}^{2}\left(\cos(\frac{x_i}{\sqrt{i}})\right)$	MN	30	[-600,600]
F5 (Powell)	$\sum_{i=1}^{n/k}(x_{4i-3}+10x_{4i-2})^2 + 5(x_{4i-1}-x_{4i})^2 + (x_{4i-2}-x_{4i-1})^4 + 10(x_{4i-3}-x_{4i})^4$	UN	24	[-4,5]
F6 (Rastrigin)	$\sum_{i=1}^{D}(x_i^2 - 10\cos(2\pi x_i) + 10)$	MS	30	[-5.12,5.12]
F7 (Schwefel)	$\sum_{i=1}^{n} -x_i \sin(\sqrt{\lvert x_i \rvert})$	MS	30	[-500,500]
F8 (Schwefel 1.2)	$\sum_{i=1}^{D}\left(\sum_{j=1}^{i} x_j\right)^2$	UN	30	[-100,100]
F9 (SumSquares)	$\sum_{i=1}^{D} i^2 x_i^2$	US	30	[-10,10]

The simulation results are compared with results of Cuckoo Search [21], a modernized implementation of PSO, which is known as PSO-2007 [22], Differential Evolution (DE) [23] and Artificial Bee Colony (ABC) algorithm [8]. We used 9 benchmark functions in [21] in order to test the performance of the CK, PSO, DE,

Table 2. Experimental results with CS/PSO

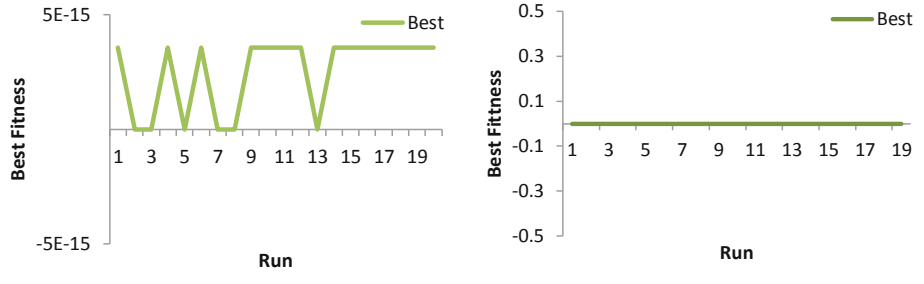
		CK	PSO	DE	ABC	CS/PSO
F1	Min	4.40E-15	8.00E-15	4.40E-15	2.22E-14	0.00E+00
	Mean	4.40E-15	8.00E-15	4.40E-15	3.00E-14	3.55E-15
	StdDev	0.00E+00	0.00E+00	0.00E+00	2.50E-15	1.67E-15
F2	Min	6.67E-01	6.67E-01	6.67E-01	1.40E-15	6.67E-01
	Mean	6.67E-01	3.27E+01	6.67E-01	2.30E-15	6.67E-01
	StdDev	5.00E-16	9.86E+01	2.00E-16	5.00E-16	2.42E-16
F3	Min	-1.00E+00	-1.00E+00	-1.00E+00	-1.00E+00	-1.00E+00
	Mean	-3.00E-01	-1.00E+00	-1.00E+00	-1.00E+00	-1.00E+00
	StdDev	4.70E-01	0.00E+00	0.00E+00	0.00E+00	0.00E+00
F4	Min	0.00E+00	0.00E+00	0.00E+00	0.00E+00	0.00E+00
	Mean	0.00E+00	9.23E-03	1.11E-03	0.00E+00	0.00E+00
	StdDev	0.00E+00	1.06E-02	3.44E-03	0.00E+00	0.00E+00
F5	Min	7.98E-08	2.11E-05	0.00E+00	2.72E-04	4.80E-13
	Mean	1.51E-07	5.18E-05	0.00E+00	4.49E-04	3.39E-11
	StdDev	5.45E-08	2.16E-05	0.00E+00	6.66E-05	5.17E-11
F6	Min	3.81E-04	1.39E+01	7.96E+00	0.00E+00	0.00E+00
	Mean	1.28E+00	2.80E+01	1.54E+01	0.00E+00	0.00E+00
	StdDev	1.04E+00	7.87E+00	4.96E+00	0.00E+00	0.00E+00
F7	Min	-1.26E+04	-1.04E+04	-1.22E+04	-1.26E+04	-8.84E+03
	Mean	-1.25E+04	-8.93E+03	-1.19E+04	-1.26E+04	-7.33E+03
	StdDev	7.64E+01	9.12E+02	2.04E+02	1.87E-12	7.73E+02
F8	Min	0.00E+00	1.28E-13	0.00E+00	1.08E+01	0.00E+00
	Mean	0.00E+00	7.39E-10	0.00E+00	4.75E+01	0.00E+00
	StdDev	0.00E+00	2.43E-09	0.00E+00	2.32E+01	0.00E+00
F9	Min	0.00E+00	0.00E+00	0.00E+00	3.00E-16	0.00E+00
	Mean	0.00E+00	0.00E+00	0.00E+00	4.00E-16	0.00E+00
	StdDev	0.00E+00	0.00E+00	0.00E+00	1.00E-16	0.00E+00

ABC and CS/PSO algorithms. The best values, the mean best values and standard deviation are given in table 2. Convergence diagram for functions F3 and F5 are presented in Figure 2. Figure 3 shows the stability diagrams for F1 and F8. In order to make comparison coherently, the global minimum values below 10^{-16} are assumed to be 0 in all experiments. 20 runs of algorithm were needed for each function and the maximum evaluation number was 2,000,000 for all functions. The parameter settings of CS/PSO are described as follows: The population size is set to 50 and the probability of discovery, P_a is set to 0.25. Acceleration constants $C1$, $C2$ are set to 1.5 and the inertia weight is 0.7. Settings of the algorithms CK, PSO, DE and ABC can be found in [21].



(a) (b)

Fig. 2. Convergence diagram for functions: (a) *F3* and (b) *F5*



(a) (b)

Fig. 3. Stability diagram for functions: (a) *F1* and (b) *F8*

7 Conclusion and Future Works

In this paper we combined two nature inspired algorithms and introduced the CS/PSO algorithm. With a profound look into the cuckoo birds' life style, we can observe the standard CS algorithm can be extended. In proposed algorithm we added swarm intelligence to the cuckoo birds, in order to increase the chance of their eggs survival. By the use of swarm intelligence it can be seen the hybrid algorithm observe more search space and can effectively reach to better solutions.

CS/PSO was evaluated on some benchmark functions and the result show that the proposed hybrid algorithm outperforms the CK with 80%, PSO-2007 with 85%, DE with 60% and ABC with more than 65% of the total time in comparison.

Our future research would include the hybridization of the CS algorithm with other nature inspired algorithms. Also the future works can include the parallelization of the cuckoo search algorithm to provide considerable gains in term of performance.

References

1. Yang, X.S.: Engineering Optimization: An Introduction with Metaheuristic Applications. Wiley Publishing, New Jersey (2010)
2. Coello Coello, C.A., Dhaenens, C., Jourdan, L. (eds.): Advances in Multi-Objective Nature Inspired Computing. SCI, vol. 272. Springer, Heidelberg (2010)
3. Yang, X.-S.: Metaheuristic Optimization: Algorithm Analysis and Open Problems. In: Pardalos, P.M., Rebennack, S. (eds.) SEA 2011. LNCS, vol. 6630, pp. 21–32. Springer, Heidelberg (2011)
4. Holland, J.H.: Adoption in Natural and Artificial Systems. University of Michigan, Ann Arbor (1975)
5. Geem, Z.W., Kim, J.H., Loganathan, G.V.: A new heuristic optimization algorithm: harmony search. Simulation 76, 60–68 (2001)
6. Dorigo, M., Di Caro, G.: The ant colony optimization meta-heuristic. McGraw-Hill, England (1999)
7. 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)
8. Karaboga, D.: An Idea Based on Honey Bee Swarm for Numerical Optimization. Technical Report-TR06, Erciyes University, Engineering Faculty, Computer Engineering Department (2005)
9. Yang, X.S., Deb, S.: Cuckoo Search via Lévy Flights. In: Proceedings of World Congress on Nature & Biologically Inspired Computing, pp. 210–214. IEEE Press, Coimbatore (2009)
10. Kennedy, J., Eberhart, R.C.: Particle swarm optimization. In: Proceedings of IEEE International Conference on Neural Networks, Piscataway, pp. 1942–1948 (1995)
11. Puranik, P., Bajaj, P., Abraham, A., Palsodkar, P., Deshmukh, A.: Human Perception-based Color Image Segmentation Using Comprehensive Learning Particle Swarm Optimization. Journal of Information Hiding and Multimedia Signal Processing 2(3), 227–235 (2011)
12. Chang, F.C., Huang, H.-C.: A Refactoring Method for Cache-Efficient Swarm Intelligence Algorithms. Information Sciences, doi:10.1016/j.ins.2010.02.025
13. Layeb, A.: A novel quantum inspired cuckoo search for knapsack problems. International Journal of Bio-Inspired Computation 3, 297–305 (2011)
14. Tuba, M., Subotic, M., Stanarevic, N.: Modified cuckoo search algorithm for unconstrained optimization problems. In: Proceedings of the 5th European Conference on European Computing Conference, pp. 263–268. WSEAS, Wisconsin (2011)
15. Walton, S., Hassan, O., Morgan, K., Brown, M.R.: Modified cuckoo search: A New Gradient Free Optimisation Algorithm. Chaos, Solitons& Fractals 44, 710–718 (2011)

16. Valian, E., Mohanna, S., Tavakoli, S.: Improved Cuckoo Search Algorithm for Feed forward Neural Network Training. Int. J. Articial Intelligence and Applications 2, 36–43 (2011)
17. Yang, X.S., Deb, S.: Engineering Optimisation by Cuckoo Search. Int. J. Mathematical Modelling and Numerical Optimisation 1, 330–334 (2010)
18. Yang, X.: Nature-Inspired Metaheuristic Algorithms, 2nd edn. Luniver Press (2010)
19. Nedjah, N., Mourelle, L.M.: Swarm Intelligent Systems. Springer, New York (2006)
20. Karaboga, D., Akay, B.: A comparative study of Artificial Bee Colony algorithm. Applied Mathematics and Computation 214, 108–132 (2009)
21. Civicioglu, P., Besdok, E.: A Conceptual Comparison of the Cuckoo Search, Particle Swarm Optimization, Differential Evolution and Artificial Bee Colony Algorithms. Artificial Intelligence Review (2011), doi:10.1007/s10462-011-9276-0
22. Xin, B., Chen, J., Peng, Z., Pan, F.: An Adaptive Hybrid Optimizer Based on Particle Swarm and Differential Evolution for Global Optimization. Science China Information Science 53, 980–989 (2010)
23. Storn, R., Price, K.: Differential Evolution - A Simple and Efficient Heuristic for Global Optimization over Continuous Spaces. Journal of Global Optimization 11, 341–359 (1997)

A Hybrid ICA/PSO Algorithm by Adding Independent Countries for Large Scale Global Optimization

Amirhossein Ghodrati[1], Mohammad V. Malakooti[2], and Mansooreh Soleimani[2]

[1] Computer Engineering Department, College of Nabi Akram, Tabriz, Iran
[2] Islamic Azad University, UAE Branch, Department of Computer Engineering
{a.h.ghodrati,m.soleimani65}@gmail.com, malakooti@iau.ae

Abstract. This paper presents the hybrid approach of Imperialist Competitive Algorithm (ICA) and Particle Swarm Optimization (PSO) for global optimization. In standard ICA, there are only two types of countries: imperialists and colonies. In the proposed hybrid algorithm (ICA/PSO) we added another type of country, '*Independent'*. Independent countries do not fall into the category of empires, and are anti-imperialism. In addition, they are united and their shared goal is to get stronger in order to rescue colonies and help them join independent countries. These independent countries are aware of each other positions and make use of swarm intelligence in PSO for their own progress. Experimental results are examined with benchmark functions provided by CEC2010 Special Session on Large Scale Global Optimization (LSGO) and the results are compared with some previous LSGO algorithms, standard PSO and standard ICA.

Keywords: ICA, Global optimization, PSO, Hybrid evolutionary algorithm and Swarm intelligence.

1 Introduction

Computing optimal solutions is intractable for many optimization problems of industrial and scientific importance [1]. The complexity of the problem of interest makes it impossible to search every possible solution or combination [2]. However, due to their complexity, the use of approximation algorithms for solving them has become almost popular in the past years. Among these optimization algorithms, modern metaheuristics are becoming increasingly popular, leading to a new branch of optimization, called metaheuristic optimization. Most metaheuristic algorithms are nature-inspired [3], some of which have been proposed for optimization problems.

PSO was formulated by Kennedy and Eberhart in 1995 [4]. It is an evolutionary computation technique which is inspired by social behavior of swarms. This algorithm is the simulation of the social behavior of birds, like the choreography of a bird flock. Each individual in the population is a particle and gets a random value in the initialization. Each particle despite the position vector contains the best personal experience and a velocity vector. The particle's velocity and its best personal

J.-S. Pan, S.-M. Chen, N.T. Nguyen (Eds.): ACIIDS 2012, Part III, LNAI 7198, pp. 99–108, 2012.

experience and the global best position all together determines a particle next movement. PSO has variety of usages such as [5] and [6].

ICA algorithm has been proposed by Atashpaz-Gargari and lucas in 2007 that has been inspired by a socio-human phenomenon [7]. This algorithm like other evolutionary algorithms starts with an initial population. Each individual of a given population is considered a country. There are two types for countries: colonies and imperialists that all together form some empires. Imperialistic competition among these empires forms the basis of the ICA evolutionary algorithm. During this competition, weak empires collapse and powerful ones take possession of their colonies. Imperialistic competition hopefully converges to a state in which there exists only one empire and its colonies are in the same position and have the same cost as the imperialists [7].

ICA and PSO are nature inspired algorithms and work well for optimization problems. In this paper, we have tried to get closer to the reality and improve the ICA by adding a new category named 'independent countries'. Despite the original idea of ICA which is based on imperialism and colonies, we considered a group of independent peaceful countries. These countries are united and they communicate with each other using swarm intelligence. They are against the imperialism countries and are in competition with them. The experimental results show that we are obtaining better results. Standard ICA is discussed in section 2. Section 3 provides an overview of PSO approach. Section 4 gives a hybrid approach of ICA with PSO and Section 5 presents the detailed experimental results to compare the performance of the proposed algorithm with other algorithms.

2 Imperialist Competitive Algorithm (ICA)

Imperialist Competitive Algorithm (ICA) is a new evolutionary algorithm in the Evolutionary Computation field based on the human's socio-political evolution. This algorithm like other algorithms starts with some initial random solutions which each one is named country. Some of the best countries in the population are selected to be the imperialists and the rest form the colonies of these imperialists. Each imperialist enclose colonies based on their power. Thus, the powerful imperialists will have more colonies than the weaker ones. In an N dimensional optimization problem a country is defined as below:

$$Country = [P_1, P_2, \dots, P_N] . \tag{1}$$

The cost of each country is evaluated with the cost function f at variables $(P_1, P_2, \dots, P_N)$ as the following:

$$c_i = f(country_i) = f(P_{i1}, P_{i2}, \dots, P_{iN}) . \tag{2}$$

When the imperialists form their empire, the imperialist countries absorb their colonies towards themselves using the absorption policy. The absorption policy shown in Fig. 1 makes the main core of this algorithm and causes the countries to move towards their minimum optima. In ICA algorithm, to search different points around the imperialist, a random amount of deviation is added to the direction of colony movement towards the imperialist. In Fig. 1, this deflection angle is shown as θ, which is chosen randomly and with an uniform distribution. In Our implementation γ is ${}^{\pi}/_{4}\,(Rad)$.

$$\theta \sim U(-\gamma, \gamma)\ . \tag{3}$$

In the absorption policy, the colony moves towards the imperialist by x unit. In Fig. 1 the distance between the imperialist and colony shown by d and x is a random variable with uniform distribution. β is greater than 1 and is near to 2. Therefore, a proper choice can be $\beta = 2$.

$$x \sim U(0, \beta \times d)\ . \tag{4}$$

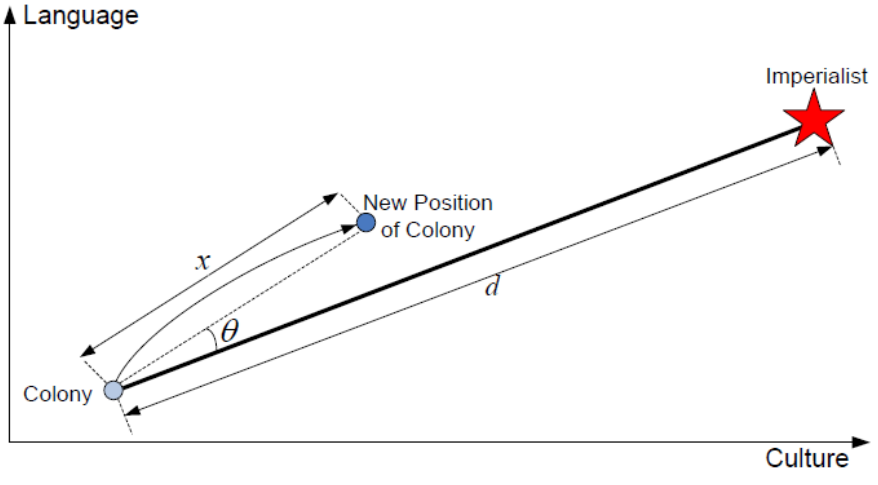


Fig. 1. Moving colonies toward their imperialist

After the absorption process, we will have the revolution operator. It is a known fact that revolution takes place in some countries, so in this algorithm revolution occurs with a probability. Revolution makes a sudden change in one or more parameters of the problem. After revolution and absorption, a colony may reach a better position, so the colony position changes according to the position of the imperialist.

For imperialistic competition first we calculate the total cost of each empire as below:

$$TC_n = cost(imperialist_n) + \xi mean\{cost(colonies\ \ of\ empire_n)\}\ . \tag{5}$$

In imperialistic competition, the weakest colony of the weakest empire will be chosen for a competition among all empires. Each empire has a chance to win that colony based on their power. Consequently, the stronger empire will have a greater chance and the weaker one will have a smaller chance.

3 Particle Swarm Optimization

Swarm Intelligence (SI) is an innovative distributed intelligent paradigm for solving optimization problems that originally took its inspiration from the biological examples by swarming, flocking and herding phenomena in vertebrates [8].

PSO is an evolutionary algorithm which is inspired by the feeding birds or fish and proposed by Kennedy and Eberhart in 1955. This algorithm like other evolutionary algorithms starts with random initial solutions and begins the process of finding the global optimum. In this algorithm, we call each solution a *particle.* Each particle moves around in the search space with a velocity. The best position explored for a particle so far is recorded and is called *pbest*. Moreover, each particle knows the best *pbest* among all the particles which is called *gbest.* By considering *pbest, gbest* and the velocity of each particle the update rule for their position is as the following equations:

$$V_{t+1} = W_t * V_t + C_1 * rand(\ \ ) * (pbest - x_t) + C_2 * rand(\ \ ) * (gbest - x_t)\ . \tag{6}$$

$$x_{t+1} = x_t + V_{t+1}\ . \tag{7}$$

Where W is inertia weight which shows the effect of previous velocity vector (V_t) on the new vector, C_1 and C_2 are acceleration constants and $rand(\ \ )$ is a random function in the range [0, 1] and x_t is current position of the particle.

4 Hybrid ICA/PSO Algorithm

In this section, we explore the details of the hybrid algorithm by adding independent countries to ICA. In standard ICA, we have colonies and imperialists only. In this paper, we have tried to get closer to the nature's reality. We have added a group of independent peaceful countries to the standard ICA. These independent countries are anti-imperialism and are united and communicate with each other. This communication is obtained from the swarm intelligence in the PSO. In the initial population, the most powerful countries are selected as independent. These countries are chosen to be powerful because the imperialists could not treat them as a colony for

themselves. In fact, these countries are anti-imperialist, so they will not constitute an empire. In each iteration of this algorithm, three more steps have been added:

4.1 Step 1. Move Independent Countries

Each country will take a new position based on three parameters:

- *pbest*: The best personal experience of that country,
- *gbest*: The best *pbest* among all the countries,
- *velocity:* The countries current velocity.

All the independent countries will move in the search space based on the following equations:

$$V_{t+1} = W_t * V_t + C_1 * rand(\ \ ) * (pbest - country_t) + C_2 * rand(\ \ ) * (gbest - country_t)\ . \tag{8}$$

$$country_{t+1} = country_t + V_{t+1}\ . \tag{9}$$

Where W is inertia weight which shows the effect of previous velocity vector (V_t) on the new vector, C_1 and C_2 are acceleration constants and $rand(\ \ )$ is a random function in the range [0, 1] and $country_t$ is current position of the country. In PSO particles move toward their best experience and global best according to equation 6 and 7, however in ICA/PSO each country update its position based on its best experience and global best of independent countries.

4.2 Step 2. Competition for Independency

As mentioned earlier, independent countries are anti-imperialism and these countries are in competition with imperialist countries. Independent countries' aim is to free the colonies from the empires and let them join independent countries to cause the collapse of all empires. In this part of algorithm, we calculate the total cost of independent countries with the mean of each ones cost.

$$TC_I = mean\{cost(independent\ countries)\}\ . \tag{10}$$

We selected the weakest colony of each empire that was weaker than independent countries. This was obtained by comparing the total cost of independent countries with the empires total cost. Then, we moved the selected colony toward the independent countries according to equation 8 and 9. This normally happens due to the fact that the colonies are not interested to be a colony, and they try to separate themselves from their empires. On the other hand, the imperialist country of that empire does not want to lose its colonies, but independent countries are more powerful than this empire and the empires are not able to stand against independent

countries. The reason for selecting the weakest colony of that empire is that empires will not attempt so much to keep the weakest colony. If the weakest colony gets stronger than its imperialist country after the movement toward independent countries, then they will leave that empire and join independent countries.

4.3 Step 3. Competition to Colonize Independent Countries

If the power of independent countries is less than that of all empires, then one of the independent countries will be available for competition between all empires like the "imperialistic competition" stage in ICA. This independent country is not interested to be a colony, but the independent countries are not powerful enough to protect that country. By 'the power of independent countries', we mean the total cost in equation 10. The pseudo-code of the ICA/PSO is presented as bellow:

```
Procedure ICA/PSO
Step 1: Initialization
   Generate some random countries;
   Select the most powerful countries as independent;
   Initialize remaining countries as empires;
Step 2: Hybrid ICA/PSO Algorithm
   Move independent countries;
   Assimilate colonies toward their imperialist;
   Countries revolution;
   Exchange imperialist with best colony if is
   necessary;
   Calculate total cost of empires;
   Competition for independency;
   Competition for colonizing independent countries;
   Imperialistic competition;
   Eliminate the powerless empires;
Step 3: Terminating Criterion Control; Repeat Step 2
```

5 Experimental Results

In this section, we show experimental results which evaluate our hybrid ICA/PSO proposed algorithm. In the field of evolutionary computation, it is common to compare different algorithms using benchmark functions, especially when the test involves function optimization. Thus, the proposed algorithm is tested on benchmark functions provided by CEC2010 Special Session on Large Scale Global Optimization [9]. This test suite is composed by five types of high dimensional problems. We picked eight functions and at least one from each type, and then we compared the results with our implementation of standard ICA, our implementation of standard PSO, jDElsgo [10], SDENS [11] and DECC-DML [12]. The mean of best values produced by the algorithms have been recorded and demonstrated for comparison in table 1. Fig. 1 shows the logarithmic scale convergence plot of ICA/PSO. Figure 2 shows the stability diagram after 3.00E+06 function evaluation for functions F1 and F6. The selected eight functions are as follows:

Table 1. Experimental results with ICA/PSO

FEs	Alg.	F1	F2	F3	F4	F5	F6	F7	F8
1.20E+05	ICA/PSO	3.15E+03	1.69E+13	9.68E+05	3.92E+02	4.5E+09	1.66E+01	2.2E+09	1.85E+06
	ICA	2.30E+04	5.48E+14	1.60E+07	4.22E+02	5.61E+12	2.14E+01	2.35E+11	2.86E+07
	PSO	2.36E+04	4.06E+13	6.24E+06	4.29E+02	9.92E+11	2.13E+01	3.74E+10	6.34E+06
	jDElsgo	1.09E+04	1.40E+14	3.15E+06	4.17E+02	7.99E+10	1.87E+01	1.64E+10	4.85E+06
	SDENS	1.19E+04	5.10E+13	2.95E+06	4.15E+02	2.61E+11	2.01E+01	1.56E+10	4.31E+06
	DECC-DML	5.75E+03	6.76E+13	4.70E+06	3.75E+02	3.84E+09	9.51E+00	4.89E+09	8.81E+06
6.00E+05	ICA/PSO	1.63E+02	4.38E+12	2.69E+05	3.88E+02	4.65E+03	1.38E-01	3.1E+08	6.46E+05
	ICA	2.24E+04	5.45E+14	1.60E+07	4.21E+02	5.61E+12	2.14E+01	2.35E+11	2.86E+07
	PSO	2.36E+04	1.31E+13	2.10E+06	4.25E+02	1.00E+11	2.05E+01	8.82E+09	3.81E+06
	jDElsgo	3.95E+03	1.39E+13	9.39E+05	2.99E+02	1.01E+06	1.22E+00	1.66E+09	1.95E+06
	SDENS	7.09E+03	1.72E+13	1.32E+06	4.13E+02	2.69E+08	6.12E+00	2.23E+09	2.07E+06
	DECC-DML	2.64E+03	1.61E+13	4.19E+06	4.47E+01	1.69E+03	1.81E-02	3.73E+08	7.27E+06
3.00E+06	ICA/PSO	1.17E-13	1.25E+12	2.52E+04	3.88E+02	6.79E+02	1.19E-09	5.5E+07	8.62E+04
	ICA	2.18E+04	5.45E+14	1.60E+07	4.21E+02	5.61E+12	2.14E+01	2.35E+11	2.86E+07
	PSO	2.36E+04	5.72E+12	7.37E+05	4.09E+02	8.55E+09	1.99E+01	1.87E+09	3.50E+06
	jDElsgo	1.25E-01	8.06E+10	1.21E+04	1.44E+02	1.53E+03	3.81E-12	3.11E+07	1.02E+05
	SDENS	2.21E+03	5.11E+12	4.13E+05	4.08E+02	9.90E+02	2.70E-05	5.63E+08	1.08E+06
	DECC-DML	2.17E+02	3.58E+12	3.80E+06	5.08E-02	9.91E+02	1.18E-13	5.92E+07	6.54E+06

- The first function $F1$ is Shifted Rastrigin's Function. The function is multimodal, shifted, separable and scalable.
- The second function $F2$ is Single-group Shifted and m-rotated Elliptic Function. The function is unimodal, shifted, single-group m-rotated and single-group m-nonseparable.
- The third function ($F3$) is $\frac{D}{2m}$ - group Shifted m-dimensional Schwefel's Problem. This Schwefel's problem is unimodal, shifted and $\frac{D}{2m}$ - group m-nonseparable.
- The forth function ($F4$) is $\frac{D}{m}$ - group Shifted and m-rotated Ackley's Function. The function is multimodal, shifted, $\frac{D}{m}$ - group m-rotated and $\frac{D}{m}$ - group m-nonseparable.
- The fifth function ($F5$) is Shifted Rosenbrock's Function. The function is multimodal, shifted and fully-nonseparable.
- The sixth function ($F6$) is Shifted Ackley's Function. The function is multimodal, shifted, separable and scalable.
- The seventh function ($F7$) is $\frac{D}{2m}$ - group Shifted m-rotated Elliptic Function. The function is unimodal, shifted, $\frac{D}{2m}$ - group m- rotated and $\frac{D}{2m}$ - group m-nonseparable.
- The eighth function ($F8$) is $\frac{D}{m}$ - group Shifted m-dimensional Schwefel's Problem. This Schwefel's problem is unimodal, shifted and $\frac{D}{m}$ - group m-nonseparable.

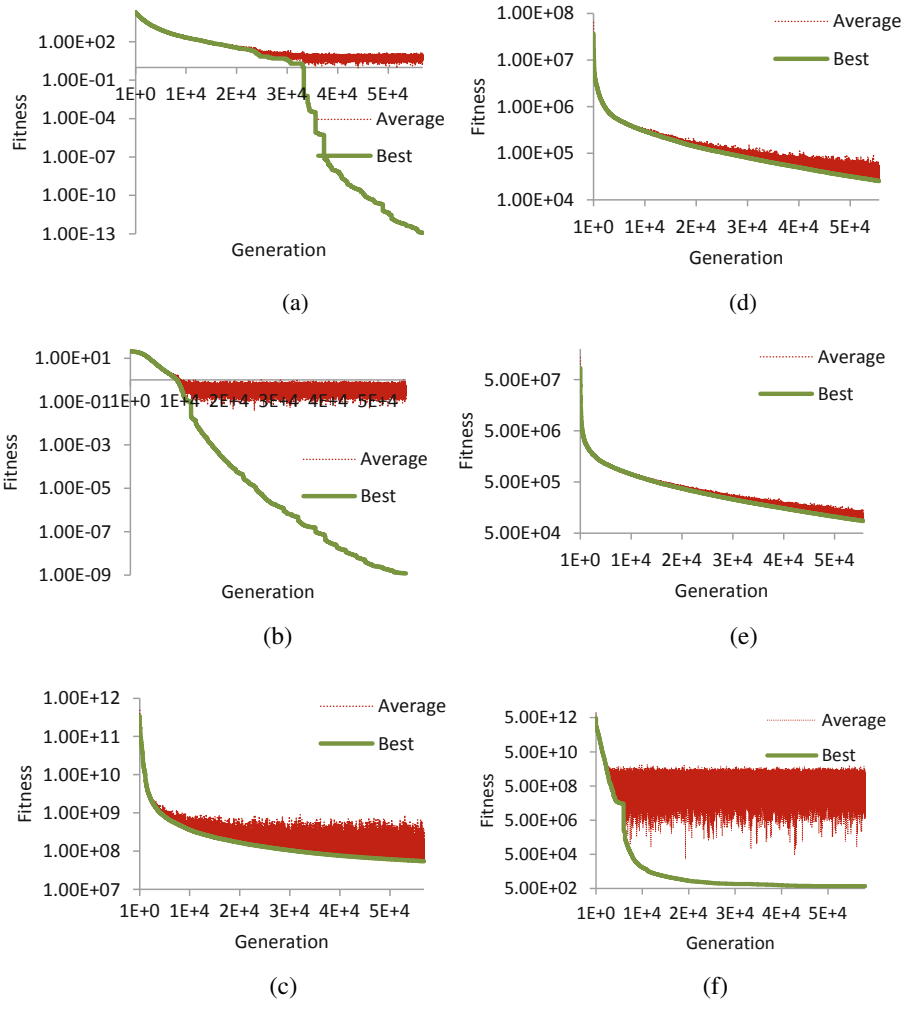


Fig. 2. Convergence diagram for functions: (a) *F1*; (b) *F3*; (c) *F5*; (d) *F6*; (e) *F7* and (f) *F8*

The control parameter used to define the degree of separability of a given function, in the given test suite is set as $m = 50$. The parameter settings of ICA/PSO are described as follows: The population size is set to 50, the number of empires is set to 5 and the number of independent countries is 10. Acceleration constants $C1$, $C2$ are set to 1.5 and the inertia weight is 0.7. The maximum number of Functions Evaluations (FEs) is set to 3e+6 for all test functions, and the study has been carried out with dimension D=1000 and 20 runs of algorithm were needed for each function.

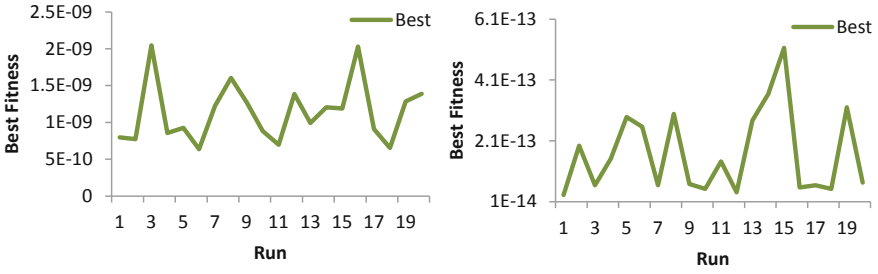


Fig. 3. Stability diagram after 3.00E+06 function evaluation for functions: (a) *F1* and (b) *F6*

6 Conclusion

This paper introduced a new hybrid ICA/PSO algorithm. Besides the colonies and imperialists in the standard ICA, we added independent anti-imperialism countries which move in the search space as the particles do in PSO. In fact, independent countries and imperialists are in competition, and when one of them finds a better solution, after a while, all countries in the search space will converge to that group. Therefore, we have more chance to find better solutions and more spaces are observed by moving countries from one group to another. ICA/PSO algorithm was evaluated on 5 of the benchmark functions provided by CEC2010 Special Session on Large Scale Global Optimization. The results show that the proposed hybrid algorithm outperforms the standard ICA, standard PSO and SDENS with 100%, jDElsgo with more than 80% and DECC-DML with more than 65% of the total time in comparison. The results show an acceptable performance of the proposed algorithm.

References

1. Talbi, E.G.: Metaheuristic: from design to implementation. Wiley Publishing, Hoboken (2009)
2. Yang, X.S.: Engineering Optimization: An Introduction with Metaheuristic Applications. Wiley Publishing, New Jersey (2010)
3. Yang, X.-S.: Metaheuristic Optimization: Algorithm Analysis and Open Problems. In: Pardalos, P.M., Rebennack, S. (eds.) SEA 2011. LNCS, vol. 6630, pp. 21–32. Springer, Heidelberg (2011)
4. Kennedy, J., Eberhart, R.C.: Particle swarm optimization. In: Proceedings of IEEE International Conference on Neural Networks, Piscataway, pp. 1942–1948 (1995)
5. Puranik, P., Bajaj, P., Abraham, A., Palsodkar, P., Deshmukh, A.: Human Perception-based Color Image Segmentation Using Comprehensive Learning Particle Swarm Optimization. Journal of Information Hiding and Multimedia Signal Processing 2(3), 227–235 (2011)
6. Chang, F.C., Huang, H.-C.: A Refactoring Method for Cache-Efficient Swarm Intelligence Algorithms. Information Sciences, doi:10.1016/j.ins.2010.02.025

7. 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)
8. Nedjah, N., Mourelle, L.M.: Swarm Intelligent Systems. Springer, New York (2006)
9. Tang, K., et al.: Benchmark Functions for the CEC 2010 Special Session and Competition on Large-Scale Global Optimization: Nature Inspired Computation and Applications Laboratory Technical report (2010), http://nical.ustc.edu.cn/cec10ss.php
10. Brest, J., Zamuda, A., Fister, I., Maučec, M.S.: Large Scale Global Optimization using Self-adaptive Differential Evolution Algorithm. In: Proceedings of IEEE Congress on Evolutionary Computation (CEC). IEEE Press, Barcelona (2010)
11. Wang, H., Wu, Z., Rahnamayan, S., Jiang, D.: Sequential DE Enhanced by Neighborhood Search for Large Scale Global Optimization. In: Proceedings of IEEE Congress on Evolutionary Computation (CEC). IEEE Press, Barcelona (2010)
12. Omidvar, M.N., Li, X., Yao, X.: Cooperative Co-evolution with Delta Grouping for Large Scale Non-separable Function Optimization. In: Proceedings of IEEE Congress onEvolutionary Computation (CEC), pp. 1762–1769. IEEE Press, Barcelona (2010)

The Modified Differential Evolution Algorithm (MDEA)

Fatemeh Ramezani[1] and Shahriar Lotfi[2]

[1] Computer Engineering Department, College of Nabi Akram, Iran
[2] Computer Science Department, University of Tabriz, Iran
f60_ramezani@yahoo.com, shahriar_lotfi@tabrizu.ac.ir

Abstract. Differential evolution (DE) is arguably one of the most powerful stochastic real-parameter optimization algorithms. DE has drawn the attention of many researchers resulting in a lot of variants of the classical algorithm with improved performance. This paper presents a new modified differential evolution algorithm for minimizing continuous space. New differential evolution operators for realizing the approach are described, and its performance is compared with several variants of differential evolution algorithms. The proposed algorithm is based on the idea of performing biased initial population. By means of an extensive testbed it is demonstrated that the new method converges faster and with more certainty than many other acclaimed differential evolution algorithms. The results indicate that the proposed algorithm is able to arrive at high quality solutions in a relatively short time limit: for the largest publicly known problem instance, a new best solution could be found.

Keywords: Modified Differential Evolution Algorithm, Differential Evolution Algorithm, Optimization, Biased Population.

1 Introduction

Continuous optimization is getting more and more attention in the last years. Many real-parameter problems from different domains can be formulated as the optimization of a continuous function. Researcher provides a variety of optimization techniques that work well in various circumstances and producing different kinds of deterministic and stochastic algorithms for optimization in the continuous domain. Consequently, there have been many studies related to real-parameter optimization.

Therefore, it is hard to devise problems for which none of the techniques can find the correct solution. Moreover, nonlinear optimization can be computationally expensive in terms of time and memory, so care must be taken when matching an algorithm to a problem.

Selecting an appropriate algorithm to solve a continuous optimization problem is not a trivial task. Although a particular algorithm can be configured to perform properly in a given scale of problems (considering the number of variables as their dimensionality), the behavior of the algorithm degrades as this dimensionality increases, even if the nature of the problem remains the same.

J.-S. Pan, S.-M. Chen, N.T. Nguyen (Eds.): ACIIDS 2012, Part III, LNAI 7198, pp. 109–118, 2012.

Several variants of DE have been proposed to improve its performance. DE is one of the most popular techniques for solving optimization problems. However, it has been observed that the convergence rate of DE do not meet the expectations in cases of highly multimodal problems. This study tries to improve DE in special high dimensional benchmark functions.

Some of the versions include Memetic Differential Evolution Algorithm [6], A Modified Differential Evolution Algorithm [7], Accelerating Differential Evolution [8], The Grouping Differential Evolution Algorithm [9], Phylogenetic Differential Evolution [10], Greedy Random Strategy [11], Preferential Mutation Operator [12], Self Adaptive DE [13], Trigonometric DE [14], Opposition Based DE [15], Neighborhood Search DE [16], Parent Centric DE [17], Modified Differential Evolution [18], Differential Evolution with Random Localization [19] etc.

In this study, a new method is presented to deal with continuous problems through different problem sizes. The proposed approach is named MDEA that stands for modified differential evolution algorithm. MDEA uses a special initialization mechanism to improve the solutions obtained by the differential evolution (DE) optimization.

In this paper we will show that the MDEA algorithm actually obtains the best results compared with several variants of DE. The experiments have been carried out on a wide range of dimensions in order to evaluate the behavior of the algorithms as the number of variables increases.

The rest of the paper is organized as follows: In Section 2 provides a brief literature overview of the DE. In Section 3, some related work is presented. In Section 4, new approach and the motivation of the DE is presented. In Section 5, the experimental results are reported.

2 Differential Evolution Algorithm (DE)

Storn and Price in 1995 [1] proposed Differential evolution. It is a relatively new optimization technique compared to evolutionary algorithms (EAs) such as Genetic Algorithms [2], Evolutionary Strategies [3], and Evolutionary Programming [4].

Like other EAs, DE is a population-based stochastic optimizer that starts to explore the search space by sampling at multiple, algorithm randomly starts with a population of NP candidate solutions: $x_{i,G}$ $(i = 1, \ldots, NP)$, where the index i denotes the population and G denotes the generation to which the population belongs [5]. DE randomly initializes a population P with K individuals (vector). DE maintains K individuals in each generation as a population. The three main operators of DE are mutation, crossover and selection.

The mutation operation of DE applies the vector differentials between the existing population members for determining both the degree and direction of perturbation applied to the individual subject of the mutation operation. The mutation process in each iteration begins for every individual $x_{i,G}$ $(i = 1, \ldots, K)$ by randomly selecting three distinct individuals $\{x_{a,G}, x_{b,G}, x_{c,G}\}$ from the population P. The i-th perturbed individual, u_i, is generated based on the three chosen individuals as follows:

$$u_i = x_{a,G} + F * (x_{b,G} - x_{c,G}) \tag{1}$$

where, $a \neq b \neq c \neq i$, F is the control parameter such that $F \in [0, 1]$. In order to form trial individual y_i, the crossover between u_i and x_i is performed. Next operation is called selection. If $f(y_i) \leq f(x_i)$ then y_i replaces x_i. The above process is repeated until termination conditions are met.

3 The Modified Differential Evolution Algorithm (MDEA)

In this Section, we describe the Modified Differential Evolution Algorithm (MDEA). The only structural difference between the proposed MDEA algorithm and the basic DE is selecting the initial candidate of solutions only. The pseudo-code of the MDEA is presented in follow:

```
Procedure MDEA
Step 0: Parameter Setting;
      Present mutation parameter F, crossover parameter Cr,
      the number of individuals in population set K, Dim set
      to problem dimension.
Step 1: Initialization;
      Initial();
      G = 0;
Step 2: While (the stopping criterion is not set)
      {
        G = G + 1;
        P'={}
        For (i = 1; i <= K; i++)
        {
          do a = ceil(rand(0,1)*K);
          while (a == i);
          do b = ceil(rand(0,1)*K);
          while (b == a||a == i);
          do c = ceil(rand(0,1)*K);
          while (c == b||b == a||a == i);
          Y_i= x_{i,G};
          u_i = x_{a,G} + F * (x_{b,G} - x_{c,G});//mutation
          jrand = ceil(Dim*rand(0,1));
          j = jrand;
          do
          (
            y_i^j = u_i^j;
            j = (j + 1) % Dim;
          }while(rand(0,1) <= Cr && j != jrand);
          P' = y_i∪P';
        }
        Selection(); /*Select NP-fittest individuals from
        P ∪ P' as next population*/
      }
```

3.1 Initialization

Classical DE initializes population within the search range randomly and uniformly. The values of parameters of a newly generated vector exceed the corresponding upper and lower bounds. This algorithm starts by generating a set of candidate solutions in the search space of the optimization problem. The generated points are called the initial population. Variables range, *[min, Max]*, divides to equal sub-range based on the population size (*NP*). The pseudo-code of initialization is presented in follow:

```
...
For (i = 1; i <= K; i++)
{
   For (j = 1; j <= Dim; j++)
   {
      x_{i,G}[j]=(i-1)*(Max-min)/(NP-1) + min;
   }
}
...
```

3.2 Selection

The DE consists of the replacement of this current population by a better fit new population. In selection operation, the fitness value of each trial vector $f(y_i)$ is compared to that of its corresponding target vector $f(x_{i,G})$ in the current population. If the trial vector has smaller or equal fitness value (for minimization problem) than the corresponding target vector, the trial vector will replace the target vector and enter the population of the next generation. Otherwise, the target vector will remain in the population for the next generation. MDEA-1 uses this strategy but MDEA-2 uses different selection operation. Each trial vectors store in P', at last selection method select number of *NP* vectors from sorted union of current and generated population, $P \cup P'$. Thus, every new population is an improvement over the earlier one.

3.3 Crossover

A suggestion leads to modify the classic DE can be found in [9]. In crossover operation, new point (u_i) is subjected to a crossover with the current individual (x_i) with a probability of crossover (C_r), yielding a candidate individual (y_i). This individual (y_i), is evaluated and if found better than x_i then it replaces x_i else x_i remains. Thus it obtains a new individual in which all individuals are either better than or as good as the current individuals. This new individual is used for the next iteration.

Exponential crossover in [9] uses a randomly selected index value (*jrand*). Corresponding parameter, u_i^j, is copied from u_i to x_i. This index causes y_i to be different from the parent individual x_i. As long as rand(0,1)$\leq C_r$, parameters continue to be taken from the u_i, since the first time rand(0,1)>Cr, the current and all remaining parameters are taken from the target individual x_i.

4 Evaluation and Experimental Results

For evaluating performance of the proposed algorithm, the simulation results are compared with results of different algorithms. The control parameters, crossover rate and scaling factor F, are fixed at 0.5 and 0.4 respectively. For dimension of 5E+02, the maximum number of evaluated fitness (FEs) allowed was set to 2E+05 and dimension of 5E+03 set to 5E+05. A total of 20 runs for each experimental setting were conducted and the average fitness of the best solutions throughout the run was recorded.

4.1 Common Benchmark Suite

In order to check the compatibility of the proposed MDEA algorithm we have tested it on a suite of nine benchmark functions; the mathematical models of the test problems are given in Table 1. Performance comparisons of MDEA algorithm is performed with classical DE on the basis of standard performance measures like average fitness functions value and number of function evaluations (FEs). From the numerical results given in Table 2, obviously MDEA algorithm gave better performance than classical DE in all the test cases. For all function, the difference in the average fitness function values for DE and MDEA is quite visible.

Table 1. Benchmark function (F1-F9)

Title	Function	Range
F1(Sphere)	$\sum_{i=1}^{D} x_i^2$	$-5.12 \le x_i \le 5.12$
F2(Rosenbrock)	$\sum_{i=1}^{D-1} 100(x_i^2 - x_{i+1})^2 + (1 - x_i)^2$	$-2.048 \le x_i \le 2.048$
F3(Rastrigin)	$\sum_{i=1}^{D} \left(x_i^2 - 10\cos(2\pi x_i) + 10\right)$	$-5.12 \le x_i \le 5.12$
F4(Griewangk)	$1 + \sum_{i=1}^{D} \left(\frac{x_i^2}{4000}\right) - \prod_{i=1}^{2} \left(\cos\left(\frac{x_i}{\sqrt{i}}\right)\right)$	$-600 \le x_i \le 600$
F5(Schwefel)	$\sum_{i=1}^{D} 418.9829 - x_i sin\left(\sqrt{\lvert x_i \rvert}\right)$	$-500 \le x_i \le 500$
F6	$\sum_{i=1}^{D} 10^{i-1} x_i^2$	$-10 \le x_i \le 10$
F7(Schwefel 1.2)	$\sum_{i=1}^{D} \left(\sum_{j=1}^{i} x_j\right)^2$	$-100 \le x, y \le 100$
F8(SumSquares)	$\sum_{i=1}^{D} i^2 x_i^2$	$-1 \le x_i \le 1$
F9(Ackley)	$-20e^{-0.2\sqrt{\frac{1}{D}\sum_{i=1}^{D} x_i^2}} - e^{\frac{1}{D}\sum_{i=1}^{D} \cos(2\pi x_i)} + 20 + e$	$-30 \le x_i \le 30$

Table 2. Comparing classical DE, MDEA-1 and MDEA-2

	Dim = 500			Dim = 5000		
	DE	MDEA-1	MDEA-2	DE	MDEA-1	MDEA-2
F1	1.02E+05	3.51E-226	**7.79E-291**	4.14E+04	**0.00E+00**	**0.00E+00**
F2	5.77E+09	**0.00E+00**	**0.00E+00**	2.35E+06	**0.00E+00**	**0.00E+00**
F3	4.33E+03	**0.00E+00**	**0.00E+00**	5.02E+04	**0.00E+00**	**0.00E+00**
F4	8.49E+02	**0.00E+00**	**0.00E+00**	1.46E+05	**0.00E+00**	**0.00E+00**
F5	1.76E+05	**6.36E-03**	**6.36E-03**	1.98E+06	**6.36E-02**	**6.36E-02**
F7	2.29E+07	2.14E-219	**2.29E-285**	-	-	**-**
F8	7.40E+05	2.59E-221	**4.11E-287**	1.34E+10	**0.00E+00**	**0.00E+00**
F9	1.32E+01	**4.44E-16**	**4.44E-16**	1.21E+01	**4.44E-16**	**4.44E-16**

Table 3 shows comparing classical DE [8], DeahcSPX [8] and MDEA-2 at dimension 30 after 3E+05 FEs. MDEA-2 performs better expect in F_{pn1}.

Table 3. Comparing classical DE, DeahcSPX and MDEA-2

F	DE[8]	DeahcSPX[8]	MDEA-2
F_{sph}	5.73E-17	1.75E-31	**0.00E+00**
F_{ros}	5.20E+01	4.52E+00	**0.00E+00**
F_{ack}	1.37E-09	2.66E-15	**4.44E-16**
F_{grw}	3.66E-03	2.07E-03	**0.00E+00**
F_{ras}	2.55E+01	2.14E+01	**0.00E+00**
F_{sch}	4.90E+02	4.71E+02	**0.00E+00**
F_{sal}	2.52E-01	1.80E-01	**4.51E-04**
F_{wht}	3.10E+02	3.06E+02	**0.00E+00**
F_{pn1}	4.56E-02	**2.07E-02**	1.05E-01
F_{pn2}	1.44E-01	1.71E-31	**1.35E-32**

Table 4 shows comparing classical DE [20], ACDE [20] and MDEA-2 at dimension 30. The maximum number of iterations allowed was set to 5000. The population size is taken as 50 for all the test problems. A total of 30 runs for each experimental setting were conducted and the average fitness of the best solutions throughout the run was recorded.

Table 4. Comparing classical DE, ACDE and MDEA-2

Function	Fitness Value			Average FEs		
	DE[20]	ACDE[20]	MDEA-2	DE	ACDE	MDEA-2
Rastringin	2.99E+01	1.33E-01	**0.00E+00**	250050	55676	7098
Sherical	6.87E-05	5.82E-05	**0.00E+00**	57000	17590	115208
Griewank	7.70E-05	6.26E-05	**0.00E+00**	175570	26120	5758
Rosenbrock	2.63E+01	3.44E+01	**0.00E+00**	250050	125320	11151
Schewefel	-12474.7	-12530	**-12569.49**	122525	44786	120090
Ackley	1.83E-04	1.72E-04	**4.44E-16**	100655	30526	120090

Table 5 shows comparing PSO [21], ABC [22] and MDEA-2.

Table 5. Comparing MDEA-2 with ABC and PSO

F	Dim.	Population	Generation	ABC	PSO	MDEA-2
	100	500	1000	8.59E-05	1.99E+00	**0**
F1	500	600	1500	2.26E+02	3.90E+03	**0**
	1000	800	2000	1.46E+03	8.00E+03	**0**
	100	500	1000	2.22E+02	3.15E+04	**0**
F2	500	600	1500	8.44E+03	1.99E+05	**0**
	1000	800	2000	5.09E+04	4.21E+05	**0**
	100	500	1000	5.39E+01	6.05E+02	**0**
F3	500	600	1500	1.93E+03	8.59E+03	**0**
	1000	800	2000	6.05E+03	1.74E+04	**0**
	100	500	1000	9.04E-03	5.24E-01	**0**
F4	500	600	1500	9.48E+02	5.10E+01	**0**
	1000	800	2000	4.86E+03	2.98E+02	**0**
	100	500	1000	1.11E-03	1.39E+02	**0**
F8	500	600	1500	5.23E+05	1.12E+07	**0**
	1000	800	2000	1.84E+08	9.50E+07	**0**
	100	500	1000	2.12E+00	3.37E+00	**0**
F9	500	600	1500	1.49E+01	2.09E+01	**0**
	1000	800	2000	1.77E+01	1.74E+04	**0**

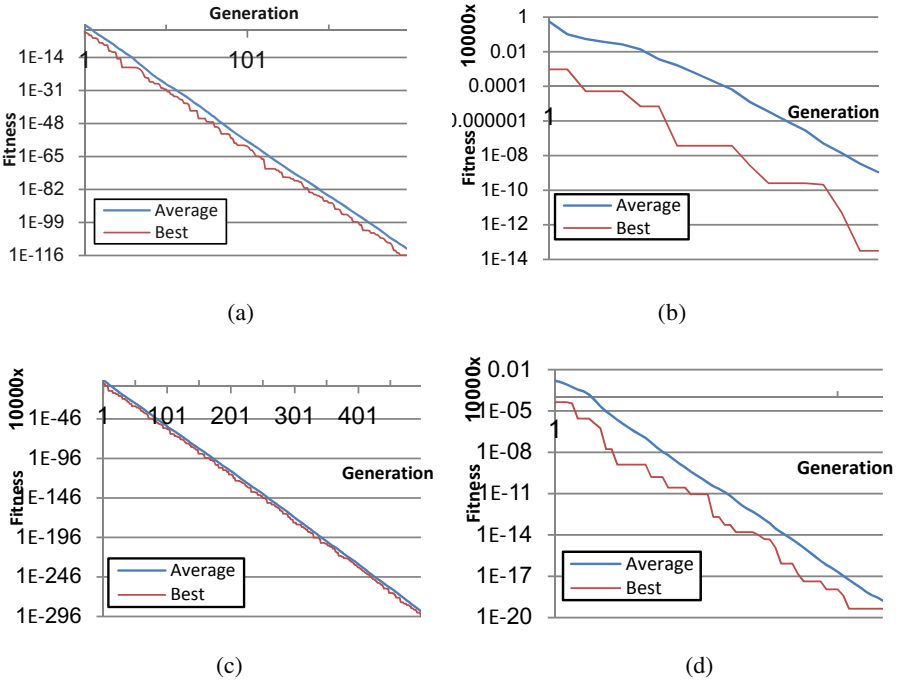


Fig. 1. Convergence diagram of MDEA (D = 500): (a) F1 (b) F2 (c) F8 (d) F9

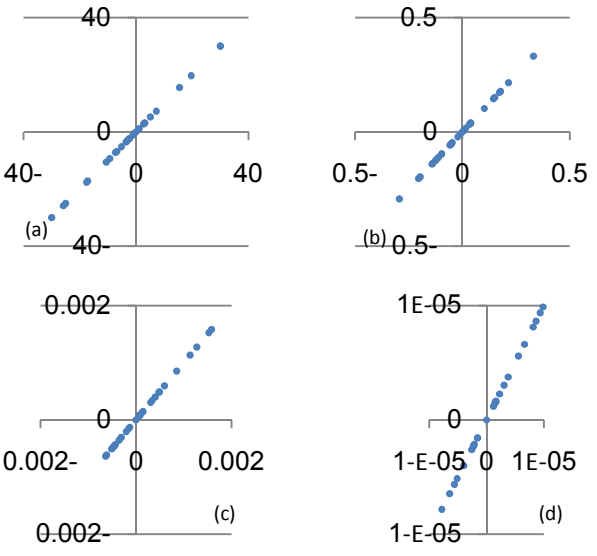


Fig. 2. Generated Points of F9 (D = 2): (a) Generation = 0 (b) Generation = 10 (c) Generation = 20 (d) Generation = 30

The convergence diagram of proposed DE algorithms for selected benchmark functions are shown in Fig. 1. Generated points for F9 in different iteration are shown in Fig. 2. The convergence diagrams are plotted with logarithmic scale for the y-axis in order to be able to capture the convergence trend over a wide range of values. In different iteration points are on the line given by the equation y = x. it illustrate modified algorithm are good for function which optimal is on this line.

4.2 Discussion

This paper proposes a novel approach based on modified initialization and its performance is evaluated using various test functions. Result shows MDEA performs better than DE [8], DEahcSPX [8], DE [20], ACDE [20], PSO and ABC in all special test functions. It shows that novel algorithm performs better.

5 Conclusion and Future Works

In this paper we proposed a modified version of basic DE called MDEA based on initialization points. The simulation of results showed that the proposed algorithm is quite competent for solving special problems of different dimensions in less number of function evaluations (FEs) without compromising with the quality of solution. The set of problems considered, though small and limited show the promising nature of

MDEA. However, the work is still in the preliminary stages and more modifications may be added to it to make it more robust. We are working to perform better initialization method for new benchmark functions.

References

1. Storn, R., Price, K.: Differential Evolution – a Simple and Efficient Adaptive Scheme for Global Optimization over Continuous Spaces, Technical Report TR-95-012, Berkeley (1995)
2. Goldberg, D.: Genetic Algorithms in Search Optimization and Machine Learning. Addison-Wesley (1989)
3. Back, T., Hoffmeister, F., Schwefel, H.: A Survey of Evolution Strategies. In: Proceedings of the Fourth International Conference on Genetic Algorithms and Their Applications, pp. 2–9 (1991)
4. Fogel, L.J.: Evolutionary Programming In Perspective: The Top-Down View. In: Computational Intelligence: Imitating Life, pp. 135–146. IEEE Press (1994)
5. Price, K.V., Storn, R.M., Lampinen, J.A.: Differential Evolution: A Practical Approach to Global Optimization. Springer, Berlin (2005)
6. Muelas, S., LaTorre, A., Pena, J.M.: A Memetic Differential Evolution Algorithm for Continuous Optimization. In: Ninth International Conference on Intelligent Systems Design and Applications, pp. 1080–1084 (2009)
7. Ali, M., Pant, M., Abraham, A.: A Modified Differential Evolution Algorithm and Its Application to Engineering Problems. In: International Conference of Soft Computing and Pattern Recognition, pp. 196–201 (2009)
8. Noman, N., Iba, H.: Accelerating Differential Evolution Using an Adaptive Local Search. IEEE Transactions on Evolutionary Computation 12(1), 107–125 (2008)
9. Piotrowski, A.P., Napirkowski, J.J.: The Grouping Differential Evolution Algorithm for Multi-Dimensional Optimization Problems. Control and Cybernetics 39(2) (2010)
10. de Melo, V.V., Vargas, D.V., Crocomo, M.K., Delbem, A.C.B.: Phylogenetic Differential Evolution. International Journal of Natural Computing Research 2(1), 21–38 (2011)
11. Bergey, P.K., Ragsdale, C.: Modified Differential Evolution: A Greedy Random Strategy for Genetic Recombination. Omega the International Journal of Management Science 33, 255–265 (2005)
12. Ali, M.M.: Differential Evolution with Preferential Crossover. European Journal of Operation Research 181, 1137–1147 (2007)
13. Salman, A., Engelbrecht, A.P., Omran, M.G.H.: Empirical Analysis of Self Adaptive Differential Evolution. European Journal of Operational Research 183, 785–804 (2007)
14. Fan, H.-Y., Lampinen, J.: A Trigonometric Mutation Operation to Differential Evolution. Journal of Global Optimization, 105–129 (2003)
15. Rahnamayan, S., Tizhoosh, H.R., Salama, M.M.A.: Opposition Based Differential Evolution. IEEE Transactions on Evolutionary Computation, 1–16 (2007)
16. Yang, Z., He, J., Yao, X.: Making a Difference to Differential Evolution. In: Advances in Metaheuristics for Hard Optimization, pp. 415–432. Springer, Heidelberg (2007)
17. Pant, M., Ali, M., Singh, V.P.: Differential Evolution with Parent Centric Crossover. In: Second UKSIM European Symposium on Computer Modeling and Simulation, pp. 141–146 (2008)
18. Babu, B.V., Angira, R.: Modified Differential Evolution (MDE) For Optimization of Non-Linear Chemical Processes. Computer and Chemical Engineering 30, 989–1002 (2006)

19. Kaelo, P., Ali, M.M.: A Numerical Study of Some Modified Differential Evolution Algorithms. European Journal of Operational Research 169, 1176–1184 (2006)
20. Thangaraj, R., Pant, M., Abraham, A.: A Simple Adaptive Differential Evolution Algorithm. In: World Congress on Nature & Biologically Inspired Computing, NaBIC 2009, pp. 457–462 (2009)
21. Kennedy, J., Eberhart, R.C.: Particle Swarm Optimization. In: Proceedings of IEEE International Conference on Neural Networks, Australia, pp. 1942–1948 (1995)
22. 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)

Quad Countries Algorithm (QCA)

M.A. Soltani-Sarvestani[1,*], Shahriar Lotfi [2], and Fatemeh Ramezani[3]

[1] Computer Engineering Departmant, College of Nabi Akram, Tabriz, Iran
[2] Computer Science Departmant, University of Tabriz, Tabriz, Iran
[3] Computer Engineering Departmant, College of Nabi Akram, Tabriz, Iran
shahriar_lotfi@tabrizu.ac.ir,
{soltani_mohammadamin,f60_ramezani}@yahoo.com

Abstract. This paper introduces an improved evolutionary algorithm based on the *Imperialist Competitive Algorithm* (ICA), called *Quad Countries Algorithm* (QCA). The Imperialist Competitive Algorithm is inspired by socio-political process of imperialistic competition in the real world and has shown its reliable performance in optimization problems. In the ICA, the countries are classified into two groups: Imperialists and Colonies. However, in the QCA, two other kinds of countries including Independent and Seeking Independence are added to the countries collection. In the ICA also the Imperialists' positions are fixed, while in the QCA Imperialists may move. The proposed algorithm was tested by well-known benchmarks, and the compared results of the QCA with results of ICA, GA [12], PSO [12], PS-EA [12] and ABC [11] show that the QCA has better performance than all mentioned algorithms. Among them, the QCA, ABC and PSO have better performance respectively in 50%, 41.66% and 8.33% of all cases.

Keywords: Optimization, Imperialist Competitive Algorithm (ICA), Independent countries, countries Seeking Independence and Quad Countries Algorithm (QCA).

1 Introduction

Evolutionary Algorithms (EA) [1, 2] are algorithms that inspire from nature and have many applications to solve NP problems in various fields of science. Some of the proposed Evolutionary Algorithms for optimization problems are: the Genetic Algorithm (GA) [2 ,3 ,4], which at first proposed by Holland, in 1962 [3] , Particle Swarm Optimization algorithm (PSO) [5] first proposed by Kennedy and Eberhart [5], in 1995. In 2007, Atashpaz and Lucas proposed an algorithm as Imperialist Competitive Algorithm (ICA) [6, 7], that has inspired from a socio-human phenomenon. Since 2007 attempts were performed in order to increase the efficiency of the ICA. Zhang, Wang and Peng proposed an approach based on the concept of small probability perturbation to enhance the movement of Colonies to Imperialist, in 2009 [8]. In 2010, Faez, Bahrami and Abdechiri, proposed a new method using the

* Corresponding author.

J.-S. Pan, S.-M. Chen, N.T. Nguyen (Eds.): ACIIDS 2012, Part III, LNAI 7198, pp. 119–129, 2012.

chaos theory to adjust the angle of Colonies movement toward the Imperialists' position (*Imperialist Competitive Algorithm using Chaos Theory for Optimization : CICA*)[9], and in other paper at the same year, they proposed another algorithm that applies the probability density function in order to adapt the angle of colonies' movement towards imperialist's position dynamically, during iterations(Adaptive Imperialist Competitive Algorithm: AICA) [10].

In the Imperialist Competitive Algorithm (ICA), there are only two different types of countries, Imperialists and Colonies, which Imperialists absorb their Colonies. While in the real world, there are some Independent countries which are neither Imperialists, nor Colonies. In the ICA, only the Colonies' movements toward Imperialists are considered while in the real world, each Imperialist moves in order to promote its political and cultural position. In the Quad Countries Algorithm (QCA), countries are divided into four categories: Imperialist, Colony, Seeking Independence and Independent that each category has its special movement compared to the others. In the Quad Countries Algorithm, like in the real world, an Imperialist will move if reaches to a better position compared to its current position.

The following part of this paper is arranged as follows. Section two describes a brief description of Imperialist Competitive Algorithm. Section three will explain the proposed algorithm. In section four, performance of algorithms will be analyzed and evaluated. In the section five, a conclusion will be presented.

2 The Imperialist Competitive Algorithm (ICA)

Imperialist Competitive Algorithm (ICA) at the first time proposed by Atashpaz and Lucas, in 2007 [6]. The ICA is a new evolutionary algorithm in the Evolutionary Computation (EC) field based on the human's socio-political evolution. The algorithm starts with an initial random population called countries, then some of best countries in the population select to be the Imperialists and the rest of them form the Colonies of these Imperialists. The number of initial population is N_{pop} including N_{col} Colonies and N_{imp} Imperialist. The Colonies divide among Imperialists. The initial number of Colonies of an Imperialist will be NC_n. The NC_n is initial number of Colonies of n^{th} Imperialist. To distribute the Colonies among Imperialists, according to the number of NC_n, they are randomly selected and assigned to the n^{th} Imperialist. The Imperialist countries absorb the Colonies towards themselves using the Absorption policy. The Absorption policy makes the main core of this algorithm and causes the countries move towards to their minimum optima. This policy is shown in Fig.1. In the Absorption policy, the Colony moves towards the Imperialist by x unit. The direction of movement is the vector from Colony to Imperialist, as shown in Fig.1. In this figure, the distance between the Imperialist and Colony is shown by d, and x is a random variable with uniform distribution. In the ICA, in order to search different points around the Imperialist, a random amount of deviation is added to the direction of Colony movement towards the Imperialist. In Fig.1, this deflection angle is shown as θ, which is chosen randomly and with a uniform distribution.

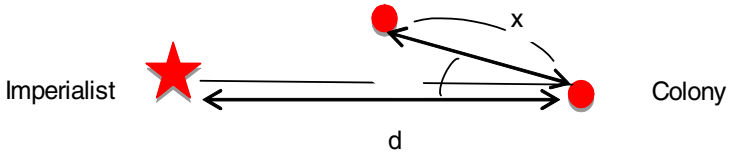


Fig. 1. Moving Colonies toward their Imperialist [6]

While Colonies moving toward the imperialist countries, a colony may reach to a better position than its imperialist, so the Colony position exchanges with position of the Imperialist.

In the ICA, the imperialistic competition has an important role. During the imperialistic competition, the weak Imperialist will lose their power and their colonies. After a while all the Imperialists except the most powerful one, will be collapse and all the Colonies will be under the control of this unique Imperialist.

3 Quad Countries Algorithm (QCA)

In this paper, a new Imperialist Competitive Algorithm is proposed which is called Quad Countries Algorithm that two new categories of countries are added to the collection of countries; Independent and Seeking Independence countries. In addition, in the new algorithm Imperialists can also move like the other countries.

3.1 Independent Country

In the real world, there are permanently countries which have been neither Colonies, nor Colonial. These Countries may perform any movements in order to take their advantages and try to improve their current situation. In the proposed algorithm, some countries are defined as Independent countries which explore search space randomly. As an illustration in figure 2, if during the search process, an Independent country achieves a better position compared to an Imperialist, they will definitely exchange their positions. The Independent countries change to a new Imperialist and will be the owner of old Imperialist's Colonies and instead of the Imperialist will changes to an Independent country and will start to explore the search space like these kinds of countries.

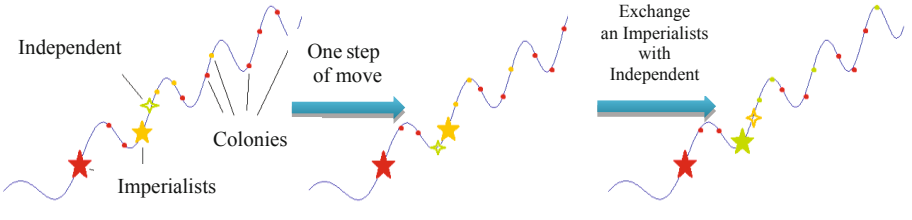
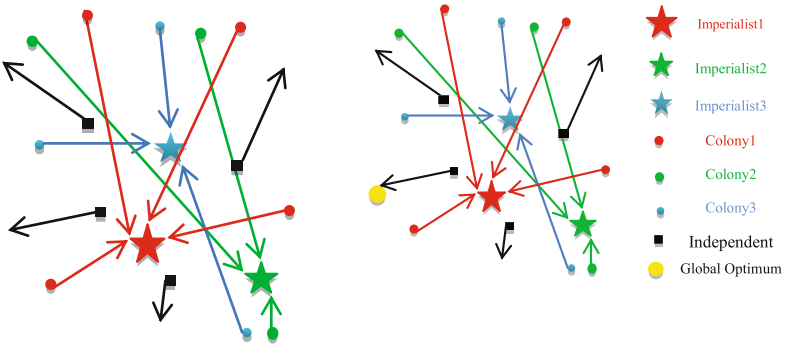


Fig. 2. Replace an Imperialist with an Independent

3.2 Seeking Independence Country

Seeking Independence Countries are countries which have challenges to the Imperialists and try to be away from them. In the main ICA, the only movement is the Colonies' movements toward Imperialists and in fact, there is only Absorption policy. While by defining the Seeking Independence Countries in proposed algorithm, there is also Repulsion policy versus Absorption policy.

Fig.3 illustrates the Repulsion Policy. As can be seen in Fig.3.a, there is only the Absorption policy that matches with the ICA. As it shows, the only use of applying Absorption policy causes that countries' positions to gets closer to each other and their surrounded space will decrease gradually and the global optima might be lost. In Fig.3.a the algorithm is converge to a local optimum. Fig.3.b illustrates the process of the proposed algorithm. The black Squares represent the Seeking Independence Countries and as can be seen, these countries can steer the search process to a direction which the other countries don't cover. It shows that, using Absorption and Repulsion policy together, will leads to better coverage of search space.



(a) Absorption policy (b) Absorption and Repulsion policy

Fig. 3. Different movement policy

To apply the Repulsion policy in the QCA, first the sum of differences between the Seeking Independence Countries and the Imperialists positions is calculated as a vector like (1) named *Center*, that is a $1\times N$ vector.

$$Center_i = \sum_{j=1}^{N_{imp}} (a_i - p_{ji}), \quad i = 1,\ 2,\ \ldots,\ N \tag{1}$$

where $Center_i$ is the sum of i^{th} component of all Imperialists, p_{ji} is i^{th} component of j^{th} Imperialist, a_i is i^{th} component of Seeking Independence Country and N indicates the problem dimensions. Then the Seeking Independence Countries will move in the direction of obtained vector as (2).

$$D = \delta \times Center, \quad \delta \in (0, \ 1) \tag{2}$$

where δ is relocation factor and D is relocation vector that its components, peer to peer sum to the Seeking Independence Country's components and obtains new position of the Seeking Independence Country.

3.3 Imperialists Movement

In the real world, all countries including Imperialists perform ongoing efforts to improve their current situation. While in the main ICA, Imperialists never move and this fixed situation sometimes leads to lose global optima or prevent to achieve better consequences. Fig.4 could be a final state of running the ICA, when only one Imperialist has remained. Since in the ICA, Imperialists have no motion, result 1 is the answer that the ICA returns. In the proposed approach, a movement, opposite to the central of gravity of its colonies is assumed for Imperialists, and the cost of this hypothetical position will be calculated. If the cost of the hypothetical position is less than the cost of the current one, the Imperialist will move to the hypothetical position, otherwise the Imperialist will not move. As can be seen in Fig.4, using this method leads to result 2 which is a better result than 1.

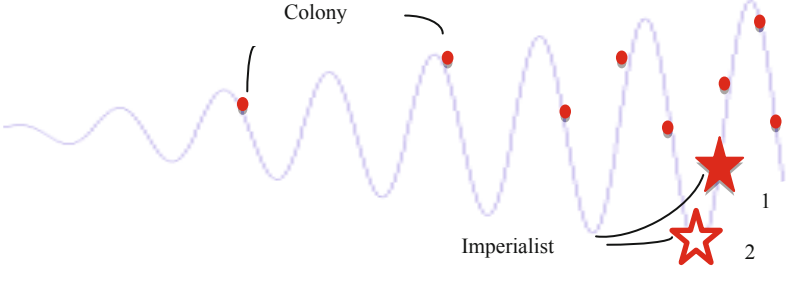


Fig. 4. A final state of ICA and QCA. Result 1 may be a final state of ICA and Result 2 may be a final state of QCA.

The movement of Imperialist is shown in equation (3).

$$\begin{aligned} &imp\_dir_i = \sum_{j=1}^{Ncol} colony_{j,i} \\ &New\_position_i = imp_i - imp + dir_i \times ieta \times rand() \\ &if\ (\cos t(New\_position_i) < \cos t(perevious\_position_i)) \\ &Then \qquad perevious\_position_i = New\_position_i \end{aligned} \tag{3}$$

Imp_dir_i is the Imperialist direction of movement of i^{th} Imperialist, $Colony_{j,i}$ is the j^{th} colony of i^{th} Imperialist, ieta is a positive value less than 1, $New\_position_i$ is hypothetical position for i^{th} Imperialist, Cost() is Cost function, and $Perevious\_position_i$ is the previous position of i^{th} Imperialist.

4 Evaluation and Experimental Results

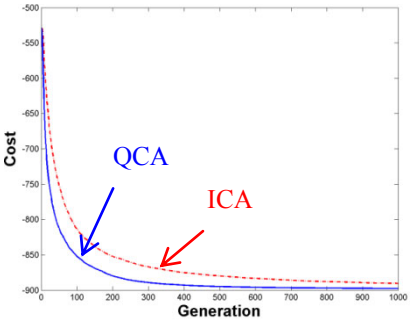
In this paper, a new algorithm based on the Imperialist Competitive Algorithm (ICA), called Quad Countries Algorithm (QCA) is introduced and was applied to some well-known benchmarks in order to verify its performance, and compare to ICA. These benchmarks functions are presented in table1.The QCA parameters set as follow: *population*=125, *ieta*=0.005, *eta*=0.01, and δ=0.01.

Each algorithm runs 100 times, and The observed results of applying the algorithms on the benchmarks are shown in table 3. *Griewank Inverse* is a hill-like function and its global optima are located on the corner of search space. Fig.5 averagely, illustrates the graph of stability and convergence of *Griewank Inverse* with 10 and 50 dimensions. It can be seen from Fig 5 that the quality of the results, the convergence of the QCA is faster than the ICA. Figures 5.c and 5.d illustrates stability graph of *Griewank Inverse* with 10 and 50 dimensions.

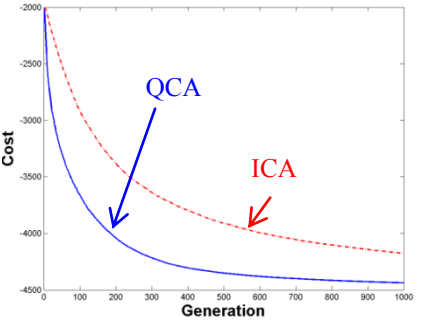
Table 1. Benchmarks for simulation

Bnechmarks	Mathematical Reprasantation	Range
Ackley	$f(x) = -20\exp\left(-0.2\sqrt{\frac{1}{n}\sum_{i=1}^{D} x_i^2}\right) - \exp\left(\frac{1}{n}\sum_{i=1}^{D}\cos 2\pi x_i\right) + 20 - e$	[-32.768, 32.768]
Griewank	$f(x) = \frac{1}{4000}\times\sum_{i=1}^{D} x_i^2 - \prod_{i=1}^{D}\cos\left(\frac{x_i}{\sqrt{i}}\right) + 1$	[-600,600]
Rastrigin	$f(x) = \sum_{i=1}^{D}\left(x_i^2 - 10\times\cos\left(2\pi x_i\right)\right) + 10D$	[-15,15]
Sphere	$f(x) = \sum_{i=1}^{D}\left(x_i^2\right)$	[-600,600]
Rosenbrock	$f(x) = \sum_{i=1}^{D-1}\left(100\times\left(x_{i-1} - x_i^2\right)^2 + \left(x_i - 1\right)^2\right)$	[-15,15]
Griewank Inverse	$f(x) = -\frac{1}{4000}\times\sum_{i=1}^{D} x_i^2 - \prod_{i=1}^{D}\cos\left(\frac{x_i}{\sqrt{i}}\right) + 1$	[-600,600]
Rastrigin Inverse	$f(x) = -\sum_{i=1}^{D}\left(x_i^2 - 10\times\cos\left(2\pi x_i\right) + 10\right)$	[-600,600]
Sphere Inverse	$f(x) = -\sum_{i=1}^{D}\left(x_i^2\right)$	[-600,600]

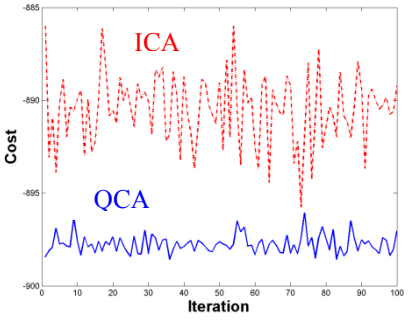
In the other comparison, the results are compared to Genetic Algorithm (GA), Particle Swarm Optimization(PSO), PS-EA and Artificial Bee Colony(ABC) in table 4. As can be seen, the results of the proposed algorithm are better than GA and PS-EA in 100 percent of all cases. But in the comparison with the QCA, the ABC and PSO, in 50 percent of cases the QCA has better performance by comparison with ABC and PSO. The ABC and PSO are 41.66 and 8.33 percent of all cases respectively, which are shown better performance.



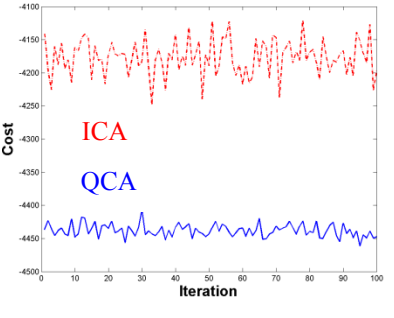
(a) the average of convergence of 100 iterations up to 1000 generation with 10 dimensions



(b) the average of convergence of 100 iterations up to 1000 generation with 50 dimensions



(c) Stability graph of 10 dimensions



(d) Stability graph of 50 dimensions

Fig. 5. The convergence and stability graphs of ICA and QCA on *Griewank Inverse*

Table 2. The result of applying benchmark on the QCA and the ICA with 2, 10, 30 and 50 dimensions

Benchmark	Alg. DIM.	Optimum	QCA			ICA			Imp.
			Best result	Results Average	SD	Best result	Results Average	SD	
Sphere	2	0	1.6384 E -26	7.4682 E -20	2.7799 E -19	2.0568 E -20	1.3710 F -10	1.1761 E -9	≈ 100%
	10	0	4.6801 E -15	1.8719 E -11	3.9881 E -11	2.5493 E -12	3.0484 E -8	6.4450 E -5	99.94%
	30	0	2.5590 E -9	7.1833 E -7	2.2583 E -6	1.0972 E -6	3.2491 E -5	3.6956 E -5	98.37%
	50	0	7.3234 E -7	3.9662 E -5	1.0098 E -4	2.6172 E -4	0.0031	0.003	99.07%
Sphere Inv	2	-7.20 E +5	-7.2 E +5	-7.2000 E +5	0.2526	-7.2000 E +5	-7.1998 E +5	14.8687	*0.003%*
	10	-3.60 E +6	-3.5995 E +6	-3.5983 E +6	783.7695	-3.5821 E +6	-3.5689 E +6	6.2142 E +3	0.82%
	30	-1.08 E +7	-1.0761 E +7	-1.0734 E +7	1.4222 E +4	-1.0506 E +7	-1.0358 E +7	5.4485 E +4	3.63%
	50	-1.80 E +7	-1.7866 E +7	-1.7755 E +7	4.0419 E +4	-1.6950 E +7	-1.6706 E +7	9.4520 E +4	6.29%
Rastrigin	2	0	0	0	0	0	1.1358 E -13	6.6520 E -13	100%
	10	0	0	1.3269 E -14	4.0851 E -14	4.4640 E -12	5.5944 E -9	1.5154 E -8	99.99%
	30	0	3.6981 E -10	1.4274 E -8	2.4467 E -8	1.6195 E -4	0.3899	0.5083	≈ 100%
	50	0	7.5566 E -7	0.0599	0.2362	1.0452	5.3211	1.7154	99.62%
Rastrigin Inv	2	-7.20 E +5	-7.2000 E +5	-7.2000 E +5	0.3104	-7.2000 E +5	-7.1999 E +5	13.0936	0.002%
	10	-3.60 E +6	-3.5995 E +6	-3.5983 E +6	906.4164	-3.5861 E +6	-3.5692 E +6	6.4272 E +3	0.82%
	30	-1.08 E +7	-1.0767 E +7	-1.0732 E +7	1.3192 E +4	-1.0486 E +7	-1.0348 E +7	4.7617 E +4	3.71%
	50	-1.80 E +7	-1.7830 E +7	-1.7757 E +7	3.4901 E +4	-1.7019 E +7	-1.6707 E +7	1.0234 E +5	6.28%
Griewank	2	0	0	0	0	0	3.7356 E-13	2.7586 E-12	100%
	10	0	8.9106 E -13	9.3103 E -9	2.2949 E -8	01.4433 E -9	6.8886 E -6	1.9415 E -5	99.86%
	30	0	4.8241 E -4	0.0144	0.0220	0.0040	0.0721	0.0522	80.03%
	50	0	0.0747	0.3832	37.9352	0.1402	0.4227	41.8484	17.2%
Griewank Inv	2	-180.0121	-179.0827	-178.9388	0.0877	-179.0774	-178.8674	0.0832	0.04%
	10	-901	-898.5777	-897.6309	0.5023	-893.9284	-890.7051	1.6002	0.79%
	30	-2701	-2688	-2681	3.6252	-2620	-2589	10.4053	3.6%
	50	-4501	-4460	-4439	9.2071	-4255	-4178	28.4462	6.28%
Ackley	2	0	8.8818 E -16	8.4754 E -13	2.0295 E -12	3.4195 E -13	5.2040 E -8	4.5546 E -7	99.99%
	10	0	1.2632 E -9	2.5552 E -8	4.8522 E -8	1.4476 E -7	2.4681 E -6	5.4074 E -6	98.99%
	30	0	1.0459 E -6	4.3273 E -6	2.5904 E -6	3.9508 E -5	1.7145 E -4	1.0189 E -4	97.54%
	50	0	3.7308 E -5	9.9126 E -5	4.1411 E -5	4.5648 E -4	0.0014	7.0191 E -4	93.54%
Schwefel	2	0	0	0	0	0	0	0	0
	10	0	0	0	0	0	0	0	0
	30	0	0	0	0	0	0	0	0
	50	0	0	0	0	0	0	0	0

Both algorithms are run 100 times, and indicate the results that are obtained after 1,000 cycle with a population having 125 individuals

Table 3. The results of GA, PSO, PS-EA, ABC, ICA and QCA

BENCH MARK	Alg DIM	GA[12]		PSO[12]		PS-EA[12]		ABC[11]		ICA		QCA	
		Mean	SD	Mean	SD	Mean	SD	Mean	SD	Mean	SD	Mean	SD
Griewank	10	0.0502	0.0295	0.0794	0.03345	0.22237	0.0781	0.00087	0.00254	6.889E-6	1.942E-5	**9.31E-9**	**2.2949E-8**
	20	1.0139	0.027	0.0306	0.02542	0.59036	0.2030	**2.01E-08**	**6.76E-08**	0.0052	0.0079	1.206E-4	1.9890E-4
	30	1.2342	0.11	0.0112	0.01422	0.8211	0.1394	**2.87E-09**	**8.45E-10**	0.0721	0.0522	0.0144	0.0220
Rastrigin	10	1.3928	0.763	2.6559	1.3896	0.43404	0.2551	**0**	**0**	5.594E-9	1.515E-8	1.327E-14	4.085E-14
	20	6.0309	1.4537	12.059	3.3216	1.8135	0.2551	1.45E-08	5.06E-08	2.154E-4	0.0016	**3.31E-11**	**6.26E-11**
	30	10.439	2.6386	32.476	6.9521	3.0527	0.9985	0.03388	0.18156	0.3899	/-4/72	**1.427E-8**	**2.4467E-8**
Ackley	10	0.5927	0.2248	**9.85E-13**	**9.62E-13**	0.19209	0.1951	7.8E-11	1.16E-09	2.468E-6	5.407E-6	2.555E-8	4.8522E-8
	20	0.924	0.226	1.178E-6	1.584E-6	0.32321	0.09735	**1.6E-11**	**1.9E-11**	3.033E-5	1.916E-5	4.311E-7	3.5441E-7
	30	1.0989	0.2496	1.491E-6	1.861E-6	0.3771	0.09876	**3E-12**	**5E-12**	1.715E-4	1.019E-4	4.327E-6	2.5904E-6
Schwefel	10	1.9519	1.3044	161.87	144.16	0.32037	1.6185	1.27E-09	4E-12	0	0	**0**	**0**
	20	7.285	2.9971	543.07	360.22	1.4984	0.84612	19.8397	45.1234	0	0	**0**	**0**
	30	13.535	4.9534	990.77	581.14	3.272	1.6185	146.857	82.3144	0	0	**0**	**0**

All algorithms indicate the results that are obtained after 500, 750 and 1,000 cycle with a population having 125 individuals

5 Conclusion and Future Works

In this paper, an improved Imperialist algorithm is introduced which is called the Quad Countries Algorithm (QCA). In the QCA, we define four categories of country including Colonial, Colony, Independent, and Seeking Independence country. Therefore, each group of countries have special motion differently compared to the others. While, in the primary ICA, there are only two categories, Colony and Colonial, and the only motion is the colonies movement toward Imperialists which is applied with absorption policy. Whereas by adding Independent countries in the QCA, a new policy which is called repulsion policy, is also added. The empirical results were found by applying the proposed algorithm to some famous benchmarks indicate that the quality of global optima solutions and the convergence speed towards the optima have remarkably increased in the proposed algorithm in comparison to the primary ICA.

Through increasing the problem dimensions, the performance of the QCA increases considerably in comparison with the ICA. Compared to the QCA and GA, PSO, PS-EA and ABC, it observed that, in 100 percent of all cases the proposed algorithm has better performance than GA and PS-EA, but in comparison with ABC and PSO, in 50 percent of cases the QCA has better performance than ABC and PSO. ABC and PSO have better performance about 41.66 and 8.33percent of cases.

Overall, the performed experiments showed that, the QCA has considerably better performance in comparison with the primary ICA and also the other evolutionary algorithms such as GA, PSO, PS-EA and ABC. The Quad Countries Algorithm (QCA) has a proper performance to solve optimization problems. However, by changing the countries' movements and defining new moving policies, its performance will increase. In fact, by define new movement policies the ability of exploration and algorithm performance will increase.

References

1. Sarimveis, H., Nikolakopoulos, A.: A Life Up Evolutionary Algorithm for Solving Nonlinear Constrained Optimization Problems. Computer & Operation Research 32(6), 1499–1514 (2005)
2. Mühlenbein, H., Schomisch, M., Born, J.: The Parallel Genetic Algorithm as Function Optimizer. In: Proceedings of The Forth International Conference on Genetic Algorithms, pp. 270–278. University of California, San diego (1991)
3. Holland, J.H.: ECHO: Explorations of Evolution in a Miniature World. In: Farmer, J.D., Doyne, J. (eds.) Proceedings of the Second Conference on Artificial Life (1990)
4. Melanie, M.: An Introduction to Genetic Algorithms. MIT Press, Massachusett's (1999)
5. Kennedy, J., Eberhart, R.C.: Particle Swarm Optimization. Proceedings of IEEE, 1942–1948 (1995)
6. Atashpaz-Gargari, E., Lucas, C.: Imperialist Competitive Algorithm: An Algorithm for Optimization Inspired by Imperialistic Competition. In: IEEE Congress on Evolutionary Computation (CEC 2007), pp. 4661–4667 (2007)

7. Atashpaz-Gargari, E., Hashemzadeh, F., Rajabioun, R., Lucas, C.: Colonial Competitive Algorithm: A novel approach for PID controller design in MIMO distillation column process. International Journal of Intelligent Computing and Cybernetics (IJICC) 1(3), 337–355 (2008)
8. Zhang, Y., Wang, Y., Peng, C.: Improved Imperialist Competitive Algorithm for Constrained Optimization. International Forum on Computer Science-Technology and Applications (2009)
9. Bahrami, H., Feaz, K., Abdechiri, M.: Imperialist Competitive Algorithm using Chaos Theory for Optimization (CICA). In: Proceedings of the 12th International Conference on Computer Modelling and Simulation (2010)
10. Bahrami, H., Feaz, K., Abdechiri, M.: Adaptive Imperialist Competitive Algorithm (AICA). In: Proceedings of The 9th IEEE International Conference on Cognitive Informatics, ICCI 2010 (2010)
11. Karaboga, D., Basturk, B.: A powerful and efficient algorithm for numerical function optimization: artificial bee colony (ABC) algorithm. Journal of Global Optimization 39(3), 459–471 (2007)
12. Srinivasan, D., Seow, T.H.: Evolutionary Computation. In: CEC 2003, Canberra, Australia, December 8-12, vol. 4, pp. 2292–2297 (2003)

An Aspectual Feature Module Based Adaptive Design Pattern for 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. Adaptability in software is the main fascinating concern for which today's software architects are really interested in providing the autonomic computing. Different programming paradigms have been introduced for enhancing the dynamic behavior of the programs. Few among them are the Aspect oriented programming (AOP) and Feature oriented programming (FOP) with both of them having the ability to modularize the crosscutting concerns, where the former is dependent on aspects ,advice and lateral one on the collaboration design and refinements. In this paper we will propose a design pattern for Autonomic Computing System which is designed with Aspect-oriented design patterns .we'll also study about the amalgamation of the Feature-oriented and Aspect-oriented software development methodology and its usage in developing a self-reconfigurable adaptive system. In this paper we used the design patterns which will satisfy the properties of an autonomic system: For monitoring we used the Observer design pattern, Decision making we used Adaptation Detector design pattern, and for Reconfiguration we used Feature-oriented Aspect insertion using participant pattern. The main objective of the system is to provide self-reconfiguring behavior at run-time by inserting into the current existing code with an aspectual feature module code without interrupting the user and to provide transparency while accessing the system. The pattern is described using a java-like notation for the classes and interfaces. A simple UML class and Sequence diagrams are depicted.

Keywords: Autonomic System, Design Patterns, Aspect-oriented Design Patterns, Feature-Oriented Programming (FOP),Aspect-Oriented Programming (AOP).

1 Introduction

The vision of autonomic computing [1] is to cut down the configuration, operational and maintenance costs of a distributed system by enabling systems to provide the self-reconfiguration, self-manage properties. So in order to achieve the vision of an autonomic computing system, it requires a system to be able to dynamically adapt to

J.-S. Pan, S.-M. Chen, N.T. Nguyen (Eds.): ACIIDS 2012, Part III, LNAI 7198, pp. 130–140, 2012.

its environment and most of the adaptations that are used in an autonomic system would tend to be crosscutting in nature. Design patterns are most often used in developing the software system to implement variable and reusable software with object oriented programming (OOP) [2]. Most of the design patterns in [2] have been successfully applied in OOPs, but at the same time developers have faced some problems like as said in [3] they observed the lack of modularity, composability and reusability in respective object oriented designs [4]. They traced this lack due to the presence of crosscutting concerns. Crosscutting concerns are the design and implementation problems that result in code tangling, scattering, and replication of code when software is decomposed along one dimension [5], e.g., the decomposition into classes and objects in OOP. To overcome this problem some advanced modularization techniques are introduced such as Aspect-oriented programming (AOP) and Feature-oriented programming (FOP).In AOP the crosscutting concerns are handled in separate modules known as aspects, and FOP is used to provide the modularization in terms of feature refinements. In our proposal of a design pattern for an autonomic computing system as developed in [6], we use the Aspect-oriented design patterns in two phases of an autonomic system i.e., for monitoring use Observer design pattern in [7] ,for decision making use adaptation detector pattern in [6].The banking application for which we apply our adaptive design pattern will consists of three more design patterns in decision-making phase 1)proxy design pattern in[8] for providing authentication facility for the user.2)participant pattern in [9] to select methods based on there characterists.3)wormhole pattern in [9]to make information from caller available to a callee . So the adaptive design pattern which we have proposed in this paper covers some of the important aspect oriented design patterns in order to provide the autonomic properties for autonomic computing system [6] see Figure 1.

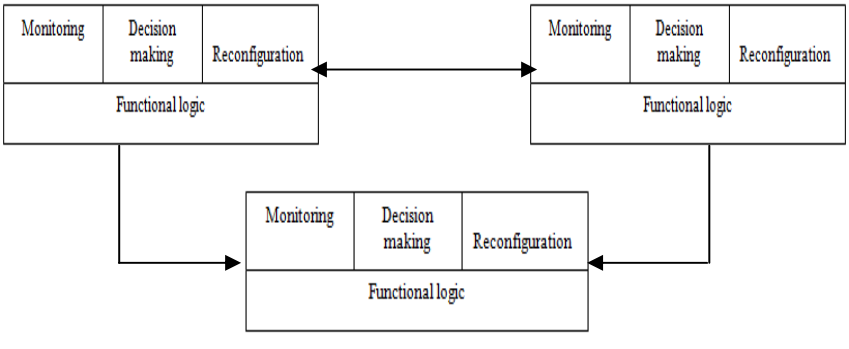


Fig. 1. Autonomic Computing

2 Related Work

In this section we present some works that deal with different autonomic systems design. There are number of publications reporting the adaptive nature of the systems

where changes occur depending upon the environments in which they are deployed. They provide the ability to monitor, to make decisions and to reconfigure at run-time. M.Vishnuvardhan and T.Ramesh paper [10] discuss applying the Adaptive Monitoring Compliance Design Pattern for autonomic systems. The authors of the paper uses adaptive design pattern called adaptive sensor factory have been proposed to make the monitoring infrastructure of the adaptive system more dynamic by fusing the sensor factory pattern, observer and strategy patterns. This pattern will determine the type of sensor that suits best for monitoring the client. In Jason O. Hallstrom and Neelam Soundarajan [7] uses observer pattern for monitoring approach for determining whether the pattern contracts used in developing a system are respected at run-time.

In Sven Apel, Thomas Leich, and Gunter Saake [11] they proposed the symbiosis of FOP and AOP and aspectual feature modules (AFMs), a programming technique that integrates feature modules and aspects. they provide a set of tools that support implementing AFMs on top of Java and C++.

3 Proposed Autonomic Design Pattern

One of the objects of this paper is to apply the Aspect-oriented design patterns to the object-oriented code in the current existing application. So that a more efficient approach to maintain the system and adapt the dynamic behaviour that is identified by the observer pattern in the autonomic system can be achieved. In the same way a more emphasis is given to include the Feature-oriented way of including or adding the new Aspect-oriented code into the current existing system. Here actually depending up on the inputs given by the client to the server, it will select the appropriate pattern to perform the operation given by the user. Then the observer pattern will monitor to check that whether the operation is executing in minimum amount of time. At this stage it will calculate time taken by the operation to execute, then it will give this as input to the decision-making pattern, which intern decides whether the task is consuming more time to execute by comparing with a threshold value, if it is true, then it will generate a Trigger informing the server that the operation have to be included in Participant pattern like a feature as a new aspect module, which takes care of time consuming operations.

4 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.[6].

4.1 Pattern Name

Aspectual Feature Module Based Adaptive Design Pattern

4.2 Classification

Structural-Monitoring

4.3 Intent

Systematically applies the Aspect-Oriented Design Patterns to an Autonomic Computing System and insertion of a new Aspect for crosscutting Time consuming operations of an application in terms of a Feature Module.

4.4 Context

Our design pattern may be used when:

- The applications/components to which we apply the AOP design patterns need to be monitored.
- For executing all the slow operations in our application in an efficient way with Aspect-oriented design patterns Implementation.
- To include the new detected slow operations as Aspectual Feature Modules [11].

4.5 Proposed Pattern Structure

A UML class diagram for the proposed design Pattern can be found in Figure 2.

4.6 Participants

(a) ClientAPI: The client will select an operation to be performed in the server container. It's a simple GUI in which he can select the operations like Authentication, Transaction Management, Credit and debit amount tasks.

(b) Server: The server is responsible for processing the client requested operations and generating the appropriate Aspect-Oriented design pattern implementation and viewing them into the main code to serve the client request.

(c) Observer: The observer is responsible for monitoring the viewing of design pattern aspect code into the application at that instance; so that it can check how efficiently the pattern is applicable for that operation execution. At the same time it will calculate the time consumed by the operation to execute and then it will pass the value to the HealthIndicator.

(d) HealthIndicator: The HealthIndicator is responsible for generating a trigger to inform the server about a slow running operation details. It will assign the task of comparing the timestamp to the analyzer.

(e) Analyzer: compares the timestamp value with a fixed threshold value to make sure that the execution of the operation will not exceed the threshold value. Then returns the result to the healthindicator.

(f) Trigger: It will inform the server with the details of the operation that has consumed more time than specified in the threshold value.

(g) CheckForSlowOpeartions: It will verify whether the operation detected is already declared as a slow operation in Participant pattern as an Aspect. If it is not the case then it will create a new aspect handler for this operation and it will make it implementable with the support of a participant pattern for the next time when a new client requests for the same operation.

4.7 Consequences

(a) This design pattern eliminates the overhead of the time consuming operations by executing them as Aspects in Participant Design Pattern.
(b) The context information from the caller to a callee can be provided directly through a wormhole pattern by avoiding linear travelling through each layer. It eliminates the need to pass the information as a set of parameters to each method in the control flow.
(c) Participant patter provides reversal of the collaboration flow. In the participant pattern, the aspect makes the class participate in the aspect collaboration, whereas in other cases, aspects affect classes without their knowledge.

4.8 Related Design Patterns

Component insertion Design Pattern [6]: This pattern can be used to safely insert and initialize a component at run time. So instead of inserting an Aspect as a Feature module we can use this pattern to insert Aspect as a component into the application.

5 Roles of Our Design Patterns in Autonomic System

Adaptation Detector: adaptation detector design pattern in [6] will interpret the monitoring data and determine when an adaptation is required. Here actually when an operation say for example, user Authentication is selected by the on-line banking application then in order to use the services of it ,the user have to authenticate himself that he is the real customer with his respective username and password. Then while doing it, the application will select a design pattern suitable at that instance (proxy design pattern [8]) and it start applying it to implement the logic of authenticating user.

Observer: The observer design pattern [7] will just monitor how much efficiently the selected design pattern have been applied with respect to the given problem and reports the adaptation detector pattern the results of how much time was consumed by

the execution of this particular operation. Then the adaptation design pattern will check to make sure that the resulted value should be equal or less than a threshold value in order to make sure that the execution of this particular operation takes less time. If not then we will make this slow operation to be handled in the participant pattern as a new Aspect Module.

Wormhole: The wormhole design pattern [9] makes information from a caller available to a callee-without having to pass the information as a set of parameters to each method in the control flow. we use this pattern in our proposed adaptive design pattern to make context information about transaction data, amount, account details to be available in a single aspect so that even if these inputs are not in single class , but we can bind them in a single aspect.

Participant: The participant design pattern [9] helps to select the methods based on their characteristics and requires the participating class to collaborate with the aspects explicitly. This pattern is used in our application to manage the execution of some slow or time consuming operations for instance consider set-amount, transaction, input/output, backup, saving operations which will obviously take time to execute. so to provide efficiency in executing such methods.

Proxy: The proxy design pattern [8] in our application helps to provide the authentication based functionality for instance, if the access to an object needs to be restricted to authorized users, a proxy can be inserted to perform access checks.

6 Feature Based Aspect Module Insertion for Reconfiguration

When the decision-making phase informs the server that an operation is executing in more amount of time than the permitted threshold level, then it's the server that checks whether the respected operation has been declared as a join point in the point cut specification of a participant pattern which will handle the slow running operations. If the condition is true then server will just ignore the information from the decision making phase, on the other hand if it is false then the operation will be declared as join point to be captured in a new aspect and this new aspect will be inserted in to the system as a new Feature-oriented module. Whenever a new client accesses the same operation it will be handled as a slow operation in participant pattern. At last the whole system will be reconfigured with this new Feature-oriented Aspect insertion method. With the programming technique, called aspectual feature modules (AFMs) proposed in [11], that implements the symbiosis of AOP and FOP is the main concept used here in our system to provide the reconfiguration property to the Autonomic Computing system.

7 Interfaces Definition for the Pattern Entities

Some of the Interfaces for the classes are provided as below:

ClientAPI

```
    Public class ClientAPI
{
String username;
String password;
Int CreditAmount;
Eventhandler()
{
}
}
```

Server

```
Public class server
{
Public void    Transactionsystem()
{
}
Public void CreditAmount()
{
}
   Protected Boolean
   setAuthentication()
{
}
   Public void
   SlowOpeartionDetails()
{
}
   Public String
   CheckforUnmarkedOperations()
{
}
Protected String AccountDetails()
{
}
}
```

Observer

```
Public class Observer
     {
String target;
String attributes;
Gettimestamp()
{
}
}
```

HealthIndicator

```
Public class HealthIndicator
{
Public  Boolean  Analyzehealth()
{
}
Public generateTrigger()
{
}
Send(Trigger t)
{
}
}
```

Analyzer

```
Public class analyzer
{
Compare(Data d,Threshold t)
{
}
}
```

Threshold

```
Public class Threshold
{
Static int Limit;
}
```

Trigger

```
Public Class Trigger
{
Source()
{
}
Timestamp()
{
}
Type() {
}}
```

The view of our proposed design pattern for the Autonomic System can be seen in the form of a class diagram see Figure 2.

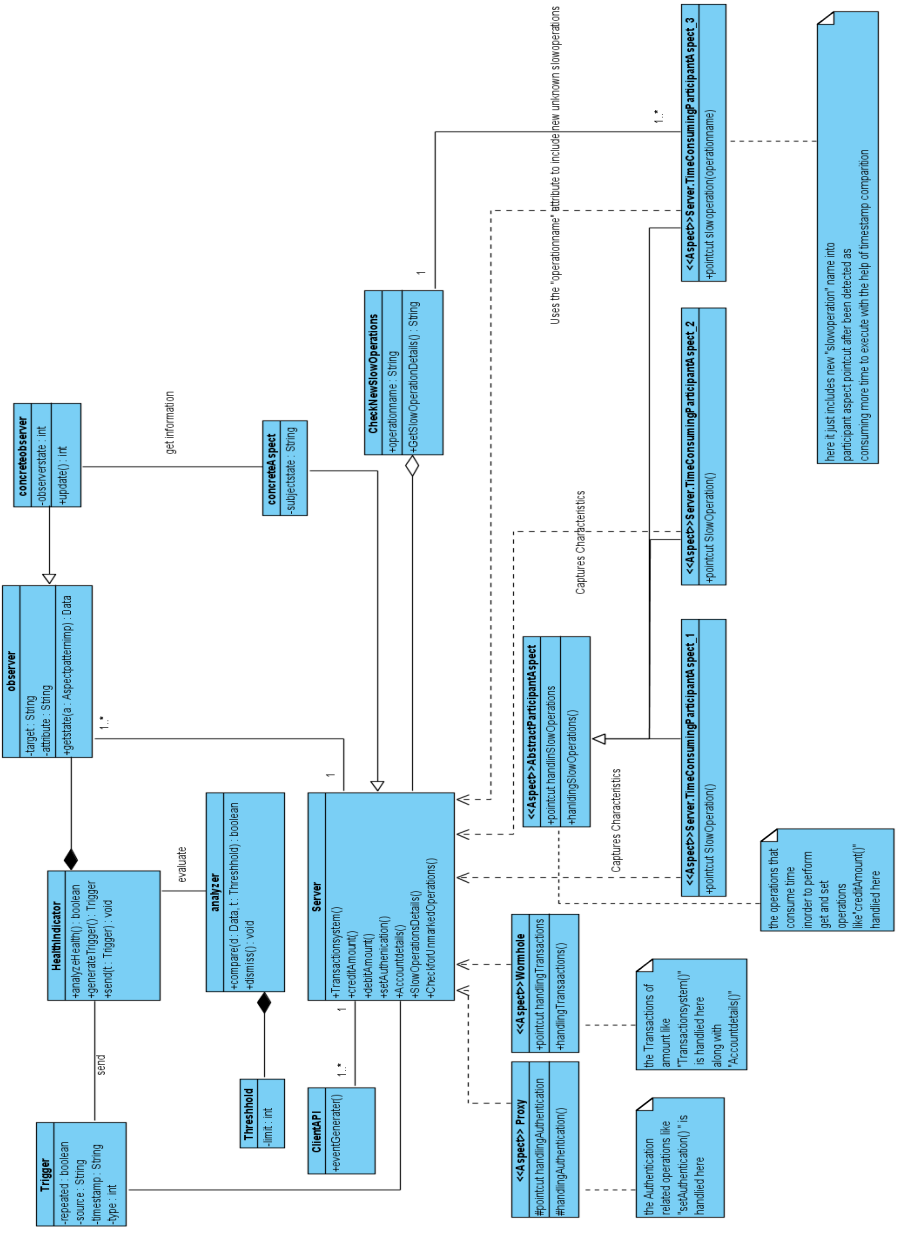


Fig. 2. Applying Design Pattern for the Autonomic System

The flow of control in the Autonomic System can be shown with a sequence diagram in Figure 3.

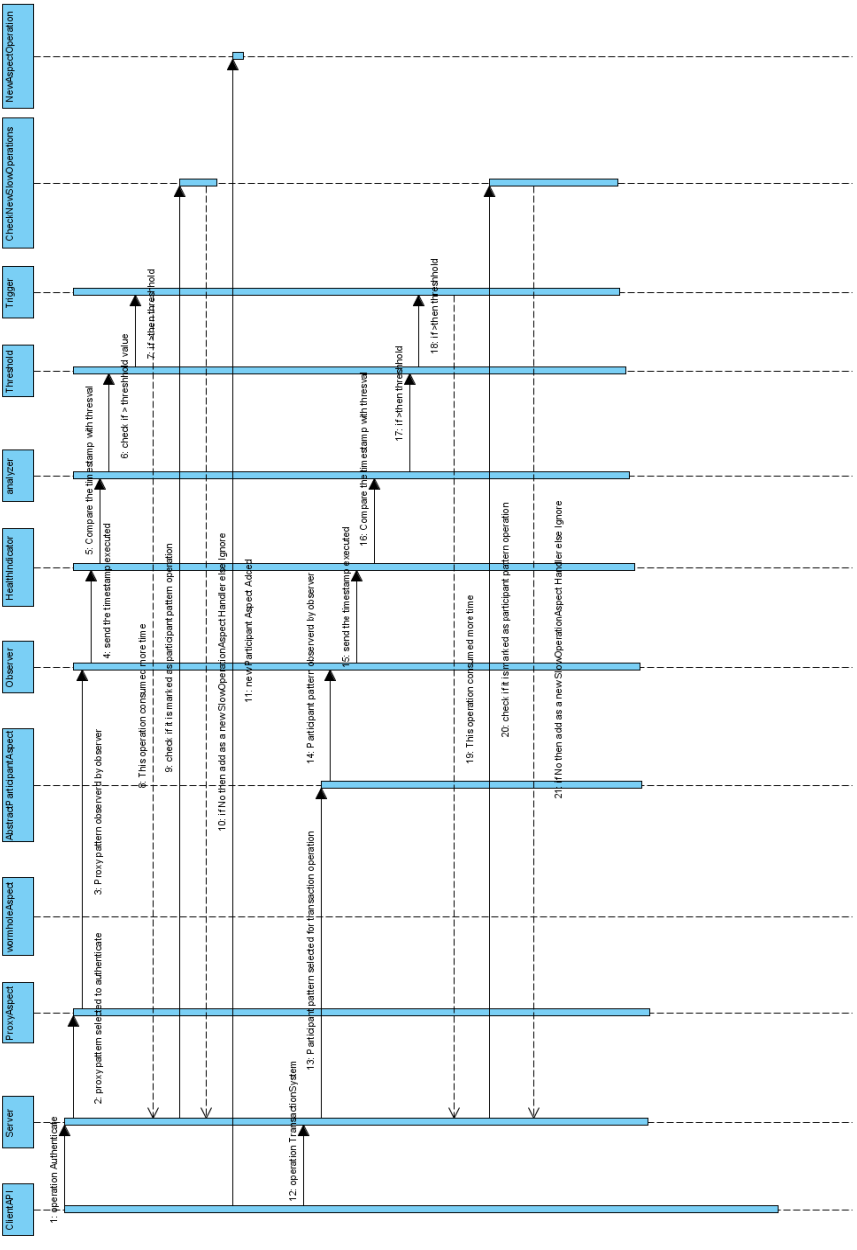


Fig. 3. Sequence Diagram for the proposed Design Pattern

8 Profiling Results

We are presenting the profiling results taken for ten runs without applying this pattern and after applying this pattern using the profiling facility available in the Netbeans IDE. The graph is plotted taking the time of execution in milliseconds on Y-axis and the run count on the X-axis. The graph has shown good results while executing the code with patterns and is shown in Figure 4.This can confirm the efficiency of the proposed pattern.

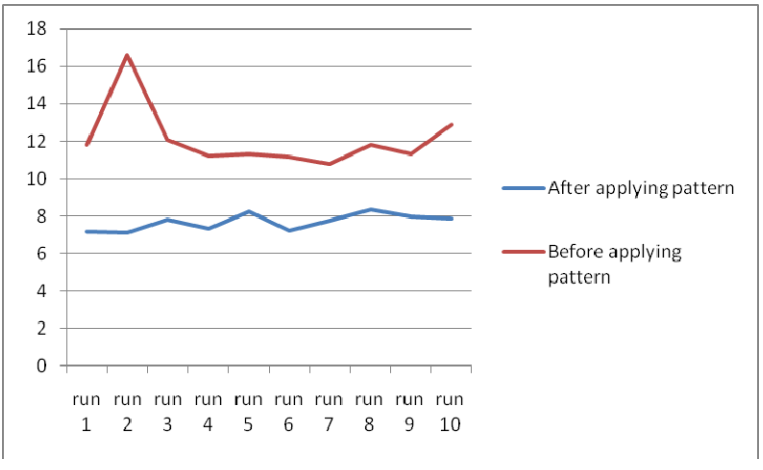


Fig. 4. Profiling data

9 Conclusion and Future Work

In this paper we have proposed a pattern to facilitate the ease of developing Autonomic computing systems that provide services to clients with the help of a set of Aspect-Oriented design patterns. So with this pattern we can handle the slow running Operations. Several future directions of work are possible. We are examining how these design patterns can be applied to the Web Service Composition through the use of aspect-oriented and feature-oriented techniques. We are also examining how the aspect-oriented design patterns can be applied in the Software product-line Architectures and also to provide the Dynamic property to an application with the amalgamation of Feature-oriented and Aspect-Oriented programming paradigms.

References

1. Soule, P.: Autonomics Development: A Domain-Specific Aspect Language Approach, pp. 14–56. Springer, Basel (2010)
2. Gamma, E., Helm, R., Johnson, R., Vlissides, J.: Design Patterns: Elements of Reusable Object-Oriented Software. Addison-Wesley (1995)

3. Kuhlemann, M., Rosenmuller, M., Apel, S., Leich, T.: On the Duality of Aspect-Oriented and Feature-Oriented Design Patterns. In: Workshop on Aspects, Components, and Patterns for Infrastructure Software (2007)
4. Hannemann, J., Kiczales, G.: Design Pattern Implementation in Java and AspectJ. In: Proceedings of the International Conference on Object-Oriented Programming, Systems, Languages, and Applications, pp. 161–173 (2002)
5. Tarr, P., Ossher, H., Harrison, W., Sutton, J.S.M.: N Degrees of Separation: Multi-Dimensional Separation of Concerns. In: Proceedings of the International Conference on Software Engineering (ICSE), pp. 107–119 (1999)
6. Ramirez, A.J.: Design Patterns for Developing Dynamically Adaptive Systems. Master's thesis, Michigan State University, East Lansing, Michigan (2008)
7. Hallstrom, J.O., Soundarajan, N., Dalton, A.R.: Monitoring Design Pattern Contracts. In: Proceedings of the Eighteenth International Conference on Software Engineering & Knowledge Engineering (SEKE 2006), San Francisco, CA, USA, July 5-7 (2006)
8. Pawlak, R., Seinturier, L., Retaillé, J.-P.: Foundations of AOP for J2EE Development, ch. 8. Apress (2005)
9. Laddad, R.: AspectJ in Action, 2nd edn., ch. 12. Manning (2010)
10. Mannava, V., Ramesh, T.: A Novel Adaptive Monitoring Compliance Design Pattern for Autonomic Computing Systems. In: Abraham, A., Lloret Mauri, J., Buford, J.F., Suzuki, J., Thampi, S.M. (eds.) ACC 2011, Part III. CCIS, vol. 190, pp. 250–259. Springer, Heidelberg (2011)
11. Apel, S., Leich, T., Saake, G.: Aspectual Feature Modules. IEEE Transactions on Software Engineering 34(2) (2008)
12. Batory, D., Sarvela, J.N., Rauschmayer, A.: Scaling Step-Wise Refinement. IEEE Transactions on Software Engineering 30(6) (2004)

Malay Anaphor and Antecedent Candidate Identification: A Proposed Solution

Noorhuzaimi Karimah Mohd Noor[1,2], Shahrul Azman Noah[2], Mohd Juzaiddin Ab Aziz[2], and Mohd Pouzi Hamzah[3]

[1] Faculty of Computer Systems & Software Engineering, Universiti Malaysia Pahang, Lebuhraya Tun Razak, 26300 Kuantan, Pahang, Malaysia
[2] Knowledge Technology Research Group, Faculty of Technology and Information Science, Universiti Kebangsaan Malaysia, 43600 Bangi, Selangor
[3] Department of Computer Science, Faculty of Science Technology, Universiti Malaysia Terengganu, 21030 Kuala Terengganu, Terengganu, Malaysia
nhuzaimi@ump.edu.my, {samn,din}@ftsm.ukm.my, mph@umt.edu.my

Abstract. This paper discusses on Malay language anaphor and antecedent candidate determination using the knowledge-poor techniques. The process to determine the candidate for anaphor and antecedent is important because the usage of pronouns in a text is not always considered as an anaphor. Sometimes pronoun referred to something outside the context or does not refer to any situation in the text. Such a situation is also exhibited in the use of pronouns in Malay language. Therefore, certain rules must be issued to identify the antecedent and anaphor candidate. Pronoun usage in Malay language does indicate the gender of the person, but to distinguish the status of the person such as imperial family, honorable people and common people. Thus, generic rules that have been used by other languages cannot simply be adapted for Malay language. The proposed solution concerns with the distance of each candidate and location of the Subject-Verb-Object (SVO) used to determine the anaphor candidate. As such, syntactic information, semantic information and distance of anaphor-antecedent are seen important to determine the antecedent candidate.

Keywords: anaphor, antecedent, Malay text, pronoun, knowledge-poor.

1 Introduction

Research on Natural Language Processing (NLP) for Malay language has started in 1977 but the focus is mainly in text pre-processing, such as morphology analyzer, stemmer and machine translation. However, there is little or none research on anaphora resolution (AR) for Malay language. Anaphora is a phenomenon that occurs in linguistic discourse level. Based on Nøklestad [1], anaphora is used to denote a number of discourses phenomena that involve referring expressions and should be solved in order to establish the coherence in a text [2]. Anaphora resolution

J.-S. Pan, S.-M. Chen, N.T. Nguyen (Eds.): ACIIDS 2012, Part III, LNAI 7198, pp. 141–151, 2012.

(AR) is a process to relate the anaphor candidate with appropriate antecedent. AR process generally involves three processes: identifying the candidate antecedents; candidates of anaphor; and the relation between appropriate antecedent and anaphor that has been selected. In this paper, determining the anaphor (third-person pronoun) and antecedent candidate are discussed and suitable rules are proposed.

Pronoun for English, Norwegian, France and Arabic language able to determine the gender of a person but in Malay language, pronouns do not differentiate the genders. As such they can be used for either gender. However, the Malay pronouns are able to distinguish the different status of the person such as imperial family, honorable persons and common persons. Thus, the needs for appropriate rules to distinguish such as status are important. This paper therefore proposed a solution for identification of anaphor and antecedent in Malay language.

2 Research Background

The anaphora resolution research has been dynamically conducted in mid 70s. This is indicated from the study by Hobbs [3] and Candace [4]. Nevertheless, research in this area gained little attention in the early 80s but bounced back at the end of the 80s. Beginning 1990s the research on anaphora resolution not only concern on the theoretical aspects but started to involve the computational linguistic field. JavaRAP [5] and Mitkov's Anaphora Resolution System (MARS) [6] are among the examples that reveal this trend.

The previously knowledge-based approach relies on hand-crafted knowledge develop by linguists. Thus, the anaphora resolution is solved by using the linguistic knowledge that represents syntax, the semantic and discourse algorithms. The preprocessing stage is conducted manually. The input is normally assumed to be perfect that has been checked or modified by the expert. Most of these systems are linguistic theoretical test (manually test by the researcher).

Hobbs [3] is one of the earlier researcher that used syntax base approach. He presented two approaches to pronoun resolution: syntax-based, known as the Naive algorithm and the semantic knowledge approach. The first approach is used for searching the noun phrase correctly with gender and number. This is done by traversing the surface parses trees of the text in order to look for an appropriate noun phrase. The second approach, on the other hand, is to identify the appropriate antecedent-anaphor using a comprehensive system for semantic analysis of English text. Hobbs defined the difficulty to solve the anaphor because of the different among the texts. He investigated the phenomena of anaphors and antecedents in all text involve and defined the candidates' sets C0, C1…CN with C0 being a subset of C1, C1 of C2 and etc. [7]. The definition of C0, C1 … CN are as shown in Table 1. The result presented [3] in shows that the algorithm work notable well.

Table 1. Candidate set prepresenation

Candidate set	Defined

C0	**If** pronoun comes before main verb **then**
	The set of entities in the current sentence and the previous sentence
	or
	if pronoun comes after main verb **then**
	The set of entities in only the current sentence

C1	The set of entities in the current sentence and the previous sentence
CN	The set of entities in the current sentence and the previous N sentences

The existent of pre-processing tools and the difficulties to adapt the knowledge-based system into another language leads the researchers to change their focus into knowledge-poor (k-poor) based system. K-poor approaches not extensively rely on linguistic or need the specific domain knowledge [8]. These systems offer the simplicity and robustness design that gives performance comparable with the knowledge-based system. Some examples that used these approaches are Mitkov's original approach (MOA) [9], RAP [10] and ARN [11]. The evaluation candidates' selection either for anaphor or antecedent is not performed but the overall valuation is implemented.

The MOA approach accepted manually POS tagged text as input [12]. Therefore, it assumes the complex syntactic, semantic and discourse analysis have been taken care of by other algorithms. All the pronouns in text assumed as anaphor. While to determine the antecedent, the noun phrases locate within a distance of two sentences before the anaphor is selected. The selected noun phrase is then inspected for gender and number agreement with the anaphor. After the agreement is agreed, the antecedent indicators that can be either a boosting or an impending capacity are applied. The boosting indicators deal with the positive score to noun phrase. On the other hand the impeding applies a negative score to a noun phrase. The two type indicator item is listed in Table 2. The approach was evaluated using a corpus of technical manuals with 223 pronouns and achieved success rate of 89.7% [9].

Table 2. Indicator item for boosting and impeding

Indicators	Item
Boosting	First noun phrases, indicating verbs, lexical reiteration, section heading preference, collocation match, immediate reference, sequential instructions, term preference
Impeding	indefiniteness, prepositional noun phrase

MARS is the new implementation of Mitkov's robust knowledge-poor approach. The differences between MARS and MOA are the use of preprocessing and how it works. MARS used the FDG-parser as its main preprocessing tool and its work fully

automatic by considering all the level of NLP processing. This version has three new indicators than MOA. Some of the indicators are different from MOA due to the availability of preprocessing tools. Three new indicators used are; boosts pronoun (pronoun can be competed as the candidate for another pronoun), syntactic parallelism (same role between antecedent and anaphor) and frequent candidate.

The processing of MARS consists of five phases [6]. The first phase all the pre-processing need for this system is implemented. The texts parse using the Conexor's FDG Parser into part of speech, morphological lemmas, syntactic functions, grammatical number, and dependency relations between tokens. The noun phrase is extracted by using the complex noun phrase extractor. The second phase, all the related anaphoric, non anaphoric and non-nominal instances (*it*) are filtered. Third phase, detection of anaphoric and the candidate antecedents are extracted based on rules. The competing candidate required to agree with the pronoun (number and gender, syntactic constraints). In forth phase, boosting and impeding factors are applied to the competing candidate. The scores were calculated in order to indicate the antecedent of the current pronouns. The last phase, the highest composite scores were select as the antecedent of the pronouns.

The approach has been evaluated using technical manual that has 2263 anaphoric pronouns and achieves the success rate 61.55% [6, 11] after the optimizations using Genetic algorithm.

The Shalom and Herbert [10] approach meant to identify the noun phrase antecedents of third-person pronouns and lexical anaphors (reflexives and reciprocals). McCord's Slot Grammar parser has been use as main preprocessing tools. The parser is use to process texts, so they are machine processable. This approach is based on salience measures derived from structure syntactic and the dynamic model in how the selection of the noun phrase as an antecedent candidate. The saliencies factor types are evaluated base on the weight that has been set by the Shalom and Herbert [10]. The saliencies factor types with the initial weights are shows in Table 3. The salience factors and anaphor binding algorithm are used to determine the candidates of antecedent.

Table 3. The saliencies factors and the initial weight

Factor Type	Description	Initial weight
Sentence recency	Current sentence, the nearest.	100
Subject emphasis	Noun phrase and pronoun in same subject	80
Existential emphasis	Predicate norminal (NP function as main predicate)	70
Accusative emphasis	Direct object	50
Indirect object and oblique complement emphasis	Indirect object and non-core argument e.g. "in the morning"	40
Head noun emphasis	NP not contain in other NP	80
Non-Adverbial emphasis	NP not contain in adverbial prepositional phrases	50

To identify the candidate of anaphor RAPS used several conditions by eliminating inappropriate anaphor. The conditions used are discussed in [10]. RAP not only defined the anaphoric pronouns but able to identify the pleonastic pronoun (PP). PP is a pronoun that does not refer to any situation in text. In this case the PP is resolved on a word *it* [10, 13]. The method to resolve the PP include lexical and syntactic test by considering a list of modal adjective.

Shalom and Herbert [10] tested the algorithm using computer manual texts and conducted a blind test on the text with 360 pronoun occurrences. The algorithm successfully identified the antecedent about 86% of the cases.

3 Identifying Malay Anaphor and Antecedent

The previous section elaborated some related work concerning ARs. Rules and patterns for identifying anaphor and antecedent have been proposed in many literatures [3, 6, 7, 10]. However, they are designed for specific language and cannot be simply adapted to suit with other languages [11]. Although some of the rules are quite similar as in the case of the morphological knowledge [14] for the English and Malay. Nonetheless, the structural patterns of the Malay language are usually dissimilar to English [15]. This section therefore discusses the issues concerning pronouns in Malay language and are they associated in determining the anaphor and antecedent.

3.1 Determination of Malay Anaphor

Pronouns in Malay language such as *dia* (he, her), *nya* (its, her, his), *baginda* (he, she) and *beliau* (he, she) are referred to human. These pronouns have been categorised into (third-person pronoun) [14]. The pronouns can be referred to both sexes either female or male. However, the uniqueness of the Malay pronouns is that they can be used to distinguish the status of persons they referred to such as honorable person, imperial family or common person.

Anaphor identification for Malay pronoun cannot be identified by looking at gender such as that proposed by Mitkov [16]. As such, it opened to the ambiguity issues and caused the process of identifying appropriate anaphor in Malay a challenging task.

Selection of appropriate anaphor in Malay pronoun has been conducted base on the location of the pronoun and the location of Subject-Verb-Object (SVO). Most of the Malay pronoun is not anaphoric within one sentence. This is quite different in English, whereby pronouns are 98% anaphoric within a sentence [17]. There are two pronoun possibilities for the anaphoric pronoun: *mereka* (they) and *nya* (its, her, and his) in the same sentence. Although, *mereka* (they) is anaphor when the location of SVO is in the second onward position. On the other hand, there is additional information to identify *mereka* as anaphor that need to be considered. The required information is the output of a tagger. On certain cases, the word *mereka* functioned as a verb in Malay text which means ‘design’. Thus tagger must be properly executed. The same thing goes with the pronoun *nya*, where the position of SVO must be considered. Only *nya* as object of the sentences or after the conjunction of target

antecedent in a sentence can be selected as an anaphor candidate. Otherwise the word *nya* refers to the outside of the context. The remaining of the Malay pronouns (i.e. *dia, ia, beliau, baginda*), based on syntactic analysis is anaphoric between two to three sentences distance. In some cases, the pronoun *nya* still anaphoric even between five sentence distances.

Table 4. Algorithms for selecting the anaphor candidate

Pronoun	

nya as human or referral other than human	**Definition 1**: tag used for human is "prnhuman" and other than human is referral. Verb that use in same sentence is represent v1 and conjunction *and* is represent as conj1
	Definition 2: SVO is subject, verb and object. SVO location is subject and object position
	
	Search for pronoun
	If pronoun= "nya" **and** tag(pronoun)= "prnhuman" **or** "referral" **then**
	If in 1st paragraph **and** 1st sentence **then**
	If location after v1 **or** after conj1 (*dan*) **or** location in S>1 or O=>1 then
	"nya" is anaphor candidate
	Else if 1st paragraph and not in 1st sentence **then**
	"nya" is anaphor candidate

Mereka	**Definition 3:** tag used to represent third-personal pronoun is "kata ganti nama"
	
	Search for pronoun
	If pronoun="mereka" **and** tag(pronoun)= "kata ganti nama" **then**
	If 1st paragraph **and** 1st sentence **then**
	If SVO location > 1 **then**
	"mereka" is anaphor
	Else if not 1st paragraph or not 1st sentence **then**
	"mereka" is anaphor

Beliau, dia, ia baginda	Search for pronoun
	If pronoun = "beliau" **or** "dia" **or** "ia" **or** "baginda" **then**
	If 1st paragraph **and** 1st sentences **then**
	If SVO location > 1 **then**
	"beliau" or "dia" or "ia" or "baginda" is anaphor
	Else not 1st paragraph **then**
	"beliau" or "dia" or "ia" or "baginda" is anaphor

Shalom and Herbert [10] identified a pronoun such as it that does not refer to any context as pleonastic. In Malay pronoun, there is a pronoun that does not refer to any situation and in any condition. The pronoun is used to emphasize something, or to convert a non-noun to noun form or known as distinctive word [14]. Several types of usage *nya* has been discussed in [18]. Karimah et.al [18] has classified the usage of *nya* into three categories for Malay AR. The first category is *nya* as human referral. Means, the pronoun *nya* is anaphoric to the human noun's phrase. The second category is *nya* that refer to other than human either event, animal, or things. The last category is *nya* that does not refer to any situation in the text either human, event, animal or things. To ensure the pronoun *nya* is the anaphor, some information must be added. The information needed is semantic information either human or referred other non-human. Table 4 shows the algorithm for selecting the anaphor previously discussed.

3.2 Determining of Malay Antecedents

The antecedent in this paper is focus on proper noun. Proper noun can be categories into organization, human, location, event, and days. The process to determine each proper noun into the selected categories is based on collocation sense. This paper will not discuss about how to categories the proper noun into its categories but to identify candidate of antecedent.

As mentioned earlier, candidates for pronoun anaphor can determine the status of the person but not the sexes. Thus, during the selection of proper noun as antecedent candidate, semantic information and syntactic information is needed. The selected proper noun will be checked for the semantic class. If the proper noun is in the semantic class of human, a set of rules are used to identify either the proper noun is selected as an antecedent or otherwise. The proposed rules are shown in Table 5 which require the searching for pronoun; location of the pronoun from the target proper noun; and the semantic class of the previous noun after the pronoun. The pronouns considered as antecedent candidate in this class are *beliau, dia, baginda, mereka* and *nya.* If the rules matched then the proper noun will be selected as the candidate for the antecedent.

If the proper noun is being semantically categorized as organization such as company, then the pronoun *nya* acts as anaphor that can determine the proper noun as antecedent candidate. In order to ensure the pronoun *nya* is referred to proper noun in this semantic class, the semantic class of the previous word that comes with a pronoun *nya* is checked and meets the rules that have been identified.

The proper noun in the semantic classes of 'location', 'thing' and 'animal' are selected as antecedent candidate based on the existent of pronoun *ia* and *nya* and meet the rules for this semantic class.

Table 5. Algorithm to determine the antecedent candidate

Categories	

SC(H) (human)	**Definition 4: pnoun** is proper noun, AC is antecedent candidate
	Definition 5: target proper noun is proper noun that aim as antecedent candidate
	Definition 6: nearest proper noun proper noun that come before target pronoun
	Definition 7: SC position is post that hold by individual
	
	Search for pronoun forward
	If Pronoun= baginda **then**
	Search phrase(**pnoun**) in imperial.txt
	If the phrase(**pnoun**)match **and** location(pronoun) = different sentence(**pnoun**) (between 2 and onward) **then**
	The target **pnoun** is AC
	If Pronoun = beliau **then**
	Search phrase(**pnoun**) in honorable.txt
	If the phrase(**pnoun**) match **and** the location(pronoun) = different sentence(**pnoun**) (between 2 and onward)**then**
	The target **pnoun** is AC
	Else If SC nearest(**pnoun** with target **pnoun**) = position **and** the location(pronoun) = different sentence(**pnoun**) (between 2 and onward) **then**
	The target pnoun is AC
	If Pronoun = nya **and** tag(pronoun)= prnhuman **then**
	If location(pronoun) = the same sentence(target **pnoun**) **and** the distance(pronoun)after v1 **or** conj1 **then**
	The target pnoun is AC
	Else If the location(pronoun) = different sentence(**pnoun**) (between 2 and onward) **then**
	The target **pnoun** is AC
	
	If Pronoun=dia **and** the location(pronoun) = different sentence(**pnoun**) (between 2 and onward) **then**
	The target **pnoun** is AC

SC(O) organization	**Definition 8: WcomeNYa** is a word that come together with pronoun *nya*, e.g. *menterinya* where *menteri* as **WcomeNYa**
	
	Search for forward pronoun
	If Pronoun=*nya* **and** tag(pronoun) = referral **then**
	If the location(pronoun) = different sentence(**pnoun**) (between 2 and onward)**and** SC **WcomeWNya**=position **then**
	The target **pnoun** is AC
	Else if location(pronoun) = the same sentence(target **pnoun**) **and** the distance(pronoun) after v1 or conj1 **and** **WcomeWNya**=position **then**
	The target **pnoun** is AC

SC(L) or SC(TG) or SM (Anl) location, thing and animal	**Definition 8:** Panil is SC for part of animal, Pthg is SC for part of thing such as door, loc is SC for location
	
	Search for forward pronoun
	If Pronoun=*nya* **and** tag(pronoun) = referral **then**
	If the location(pronoun) = different sentence(**pnoun**) (between 2 and onward) **and** SC **WcomeWNya**=Panil **or** **WcomeWNya**=Pthg **or** **WcomeWNya**=loc **then**
	The target **pnoun** is AC
	Else if location(pronoun) = the same sentence(target **pnoun**) **and** the distance(pronoun) after v1 or conj1 **and** **WcomeWNya**=Panil **or** **WcomeWNya**=Pthg **or** **WcomeWNya**=loc **Then**
	The target **pnoun** is AC
	If Pronoun=*ia* **then**
	If the location(pronoun) = different sentence with **pnoun** (between 2 and onward) **then**
	The target **pnoun** is AC

4 Discussion

There are differences between anaphor and antecedent determination in existing anaphora resolution as compared to Malay anaphora resolution. Most languages can identify an antecedent of a pronoun by the gender; the location of the antecedent and anaphor and by eliminating inappropriate anaphor. However, for Malay language status of person, semantic knowledge and tagger information is needed in order to identify the anaphor and antecedent candidate. Table 6 and 7 summarize the differences between Malay anaphora resolution and existing anaphora resolution from three parameters.

Table 6. Differences between Malay AR system and existing AR systems (anaphor selection)

Parameter	Anaphora Resolution (AR) System		
	Malay AR	RAP	MARS
Location (paragraph and sentence)	Yes	No	No
Type of anaphor	Third person pronouns	Yes	Thir person pronoun
SVO	Yes	No	No

Table 7. The differences between others AR systems (antecedent selection)

Parameter	Anaphora Resolution (AR) System		
	Malay AR	RAP	MARS
Type of antecedent	Proper noun	Proper noun and noun phrase	Proper Noun and noun phrase
Gender	No	Yes	Yes
Status (emprial, honorable, common)	Yes	No	No
Number	Yes	Yes	Yes
Head adjuct	No	Yes	Yes
Head argument	No	Yes	Yes
Distance from anaphor	Always in different sentence accept for complex sentence or some of pronoun, *nya* and *mereka*	Yes	Yes
Semantic knowledge (class)	Yes	Not known	Not known

Table 6 illustrates the parameter involved during the selecting candidate of anaphor. For Malay text, the pronoun can be considered as an anaphor candidate when the location of the pronoun is in the second sentence or in the second SVO position either as subject or object. Other AR systems, however, do not address the issued on selecting the anaphora candidates. On the other hand, Table 7 shows the elements that have been considered during the selection of antecedent candidate. Most

of the elements cannot be used to identify antecedent candidate in Malay language. This shows that the rules used by the other systems cannot be simply adapted to Malay AR problems.

5 Conclusions

As a conclusion, the process for determining the candidate for anaphor and antecedent is very much dependent on the languages. English, Norwegian and France, for example, the pronoun–antecedent can be determined via person, gender and number. However, in Malay language determining a pronoun-antecedent is by means of status and number. Such a difference is due to the fact that the pronouns in Malay language can be used for either gender. This opened to the issues of ambiguity and as such, the rules used by others [9, 10] are not directly applicable for Malay anaphor and antecedent candidate identification. Thus, semantic knowledge, tagger information, syntactic information, the distance of the anaphor-antecedent and location of SVO is needed to solve Malay anaphora and antecedent candidate selection as exhibited in this paper. Near future works involved evaluating the proposed algorithms and implementation of the rules into Malay AR system.

References

1. Nøklestad, A.: A Machine Learning Approach to Anaphora Resolution Including Named Entity Recognition, PP Attachment Disambiguation, and Animacy Detection. Faculty of Humanities, PhD, p. 298. The University of Oslo, Norway (2009)
2. Muñoz, R., Saiz-Noeda, M., Montoyo, A.: Semantic Information in Anaphora Resolution. In: Ranchhod, E., Mamede, N.J. (eds.) PorTAL 2002. LNCS (LNAI), vol. 2389, pp. 63–70. Springer, Heidelberg (2002)
3. Hobbs, J.R.: Pronoun Resolution. Research Report. City University of New York, New York (1976)
4. Candace, L.S.: Towards a Computational Theory of Definite Anaphora Comprehension in English Discourse. Massachusetts Institute of Technology (1979)
5. Qiu, L., Kan, M.-Y., Chua, T.-S.: A Public Reference Implementation of the RAP Anaphora Resolution Algorithm. In: Proceedings of the Language Resources and Evaluation Conference 2004 (LREC 2004), Lisbon, Portugal, pp. 291–294 (2004)
6. Mitkov, R., Evans, R., Orăsan, C.: A New, Fully Automatic Version of Mitkov's Knowledge-Poor Pronoun Resolution Method. In: Gelbukh, A. (ed.) CICLing 2002. LNCS, vol. 2276, pp. 168–186. Springer, Heidelberg (2002)
7. Mitkov, R.: The Past: Work in the 1960s,1970s and 1980s. In: Mitkov, R. (ed.) Anaphora Resolution, pp. 68–92. Pearson Publication, Great Britain (2002)
8. Dimitrov, M., Bontcheva, K., Cunningham, H., Maynard, D.: A Light-weight Approach to Coreference Resolution for Named Entities in Text. In: Fourth Discourse Anaphora and Anaphor Resolution Colloquium (DAARC), vol. 4. DAARC, Lisbon (2002)
9. Mitkov, R.: An approach in focus: Mitkov's robust, knowledge-poor algorithm. In: Anaphora Resolution, p. 220. Pearson Education, Great Britain (2002)
10. Shalom, L., Herbert, J.L.: An algorithm for pronominal anaphora resolution. Comput. Linguist. 20, 535–561 (1994)

11. Holen, G.I.: Automatic Anaphora Resolution for Norwegian (ARN). Department of Linguistics and Scandinavian Studies, Master, p. 142 University of Oslo, Norway (2006)
12. Ge, N.: An Approach to Anaphoric Pronouns. Computer science, PhD, p. 146. Brown University, Providence (2000)
13. Mitkov, R.: The Present: Knowledge-poor and corpus approaches in the 1990s and beyond. In: Anaphora Resolution, pp. 95–129. Pearson Education, Great Britain (2002)
14. Nik Safiah, K., Farid, M.O., Hashim, H.M., Abdul Hamid, M.: Tatabahasa Dewan. Dewan Bahasa dan Pustaka, Kuala Lumpur (2008)
15. Mohd Juzaiddin, A.A.: Pola Grammar for Automated Marking of Malay Short Answer Essay-Type Examination. Computer Science and Information Technology, PhD, p. 194. Universiti Putra Malaysia, Malaysia (2008)
16. Mitkov, R.: An integrated model for anaphora resolution. In: Proceedings of the 15th Conference on Computational Linguistics, vol. 2. Association for Computational Linguistics, Kyoto (1994)
17. Hobbs, J.: Resolving pronoun references. In: Readings in Natural Language Processing, pp. 339–352. Morgan Kaufmann Publishers Inc. (1986)
18. Karimah, M.N.N., Aziz, M.J.A., Noah, S.A.M., Hamzah, M.P.: "nya" as Anaphoric Word: A Proposed Solution. In: International Conference on Semantic Technology and Information Retrieval (STAIR), Putrajaya, pp. 249–254 (2011)

Ranking Semantic Associations between Two Entities – Extended Model

V. Viswanathan[1] and Ilango Krishnamurthi[2]

[1] Department of Computer Applications
[2] Department of Computer Science and Engineering
Sri Krishna College of Engineering and Technology, Coimbatore-641008,Tamil Nadu, India
{visuskcet,ilango.krishnamurthi}@gmail.com

Abstract. Semantic association is a set of relationships between two entities in knowledge base represented as graph paths consisting of a sequence of links. The number of relationships between entities in a knowledge base might be much greater than the number of entities. So, ranking the relationship paths is required to find the relevant relationships with respect to the user's domain of interest. In some situations, user may expect the semantic relationships with respect to specific domain closer to any one of these entities. Consider the example for finding the semantic association between the person X and person Y. If the user has already known something about the person X such as person X may be associated with financial activities or scientific research etc., then the user wants to focus on finding and ranking the relationship between two persons in which the users' context is closer to person X. In many of the existing systems, there is no consideration given into context closeness during ranking process. In this paper, we present an approach which allows the extraction of semantic associations between two entities depending on the choice of the user in which the context is closer to left or right entity. The average correlation coefficient between proposed ranking and human ranking is 0.70. We compare the results of our proposed method with other existing methods. It explains that the proposed ranking is highly correlated with human ranking. According to our experiments, the proposed system provides the highest precision rate in ranking the semantic association paths.

Keywords: Semantic Web, Semantic Association, Complex relationship, RDF, RDF Schema.

1 Introduction

Information retrieval over semantic metadata has received a great amount of interest in both industry and academia. The semantic web contains not only resources but also includes the heterogeneous relationships among them. In the current generation technologies of search engine, it is very difficult to find the relationships between entities. For example, 'how two entities are related?' is the most crucial question. Discovering relevant sequences of relationships between two entities answers this question. Semantic association represents a direct or indirect relationship between two

J.-S. Pan, S.-M. Chen, N.T. Nguyen (Eds.): ACIIDS 2012, Part III, LNAI 7198, pp. 152–162, 2012.

entities. Different entities may be related in multiple ways. For example, finding the semantic association between two persons 'X' and 'Y' who belong to film industry and they are involved in financial activities and Politics. There may be multiple paths between two entities that involve more intermediate entities that cover multiple domains. Discovering and ranking such relations based on user's interest is required. To combat this problem, Anyanwu,K et al., propose to rank semantic association [3] using six types of metrics called Subsumption (how much meaning a semantic association conveys depending on the places of its components in the RDF), Path Length (that allows preference of either immediate or distant relationships), Popularity (number of incoming and outgoing edges), Rarity (rarely occurring entity), Trust (determining how reliable a relationship is according to its origin) and context weight. In this method, path weights are calculated using these parameter values and then ranked according to their total weights. Semantic Association path with more weight is ranked first.

Suppose user knows some information about person 'X' or person 'Y' such as person 'X' involved in more financial activities or person 'Y' involved in politics etc. In such cases, user may be interested in finding that types of relationships between 'X' and 'Y' and it should be ranked first.

Consider the example for the relationships between person '**John**' and movie **'Slumdog_millionaire**' in an RDF.

John *– edit_music* ***- Human_Movie*** *– support_fund -* ***Tara_Funding_Agencies*** *– support_fund -* ***Ragam_Music*** *– member_of -* ***ARRahman*** *– provides_music -* ***Slumdog_millionaire***
John *– member_of* ***- Tara_Funding_Agency*** *- supports_fund -* ***Human_Movie*** *– associated_with -* ***Ragam_Music*** *– member_of -* ***ARRahman*** *– provides_music -* ***Slumdog_millionaire***
John *– edits_music –* ***Human_Movie*** *– associated_with -* ***Ragam_Music*** *– support_fund -****Tara_Funding_Agencies*** *– member_of -* ***ARRahman*** *– provides_music -* ***Slumdog_millionaire***

From the above example, the same set of components(properties and entities) may be scattered over the path in different possible combination. If the user is interested in Music and Finance, Anyanwu,K et al., method produce the same weights for all paths. In this method, while calculating each metric, the component values are calculated independently and then summed up. Sometimes, all the paths are having equal weights are ranked arbitrarily. In that case, user has to go through this subset of paths to find the relevant paths. Suppose user may be interested in finding and ranking relevant association according to his domain of interest which is closer to either left entity (John) or right entity (Slumdog millionaire), we have to rank these paths according to user's interest. In the existing systems, users may select the choice for favoring long path or favoring short path, favoring popularity or favoring unpopular or favoring rarity, but there are no ways to select the choice for context closeness. In the proposed method, we find and rank the semantic association paths between two entities according to the users' needs with context closeness.

The organization of this paper is as follows: In Section 2, we present an overview of the background and basic definitions of semantic association. In Section 3, an overview of some related works in the area of semantic association is given. Section 4, explains how the semantic association paths weight is evaluated and ranked. Experimental evaluation of the proposed approach is explained in Section 5. Section 6, summarizes the contribution and states the possible future work.

2 Background

The Resource Description Framework (RDF)[9][10] data model provides a framework to capture the meaning of an entity by specifying how it relates to other entities. In RDF model, concepts of entities are linked together with relations (properties). The properties are denoted by arcs and labeled with the relation name. The definition of the RDF graph is as follows:

Definition 1*(RDF graph):* RDF graph is a directed labeled graph to represent the relationship between entities.

Semantic associations are complex relationships between resource entities [4]. Most of the useful semantic associations involve some intermediate entities and relations. It helps the user to see the connection between different people, places and events. Semantic associations are based on concepts such as semantic connectivity and semantic similarity. To describe the semantic connection between two entities in domain RDF, we introduce some definition[3]:

Definition 2*(Semantic Connectivity):* Two entities e_1 and e_n are semantically connected if there exists a sequence $e_1, P_1, e_2, P_2, e_3, P_{3.....} e_{n-1}, P_{n-1}, e_n$ in an RDF graph where e_i $(1 \leq i \leq n)$ are entities and P_j $(1 \leq j < n)$ are properties.

Definition 3 *(Semantic Similarity):* Two entities e_1 and f_1 are semantically similar if there exist two semantic paths $e_1, P_1, e_2, P_2, e_3, P_{3.....} e_{n-1}, P_{n-1,} e_n$ and $f_1, Q_1, f_2, Q_2, f_3, Q_{3.....} f_{n-1}, Q_{n-1,} f_n$ semantically connecting e_1 with e_n and f_1 with f_n respectively, and that for every pair of properties P_i and Q_i, $1 \leq i < n$, either of the following conditions holds: $P_i = Q_i$ or $P_i \subseteq Qi$ or $Qi \subseteq Pi$ ($\subseteq$ means rdf:subPropertyOf), then two paths originating at e_1 and f_1, respectively, are semantically similar.

Definition 4 *(Semantic Association):* Two entities e_x and e_y are semantically associated if e_x and e_y are semantically connected or semantically similar.

3 Related Work

Several techniques have been proposed related to ranking of semantic associations. Some of them are summarized below:

For ranking the results of complex relationship searches on the semantic web, Anyanwu et al.[3] present a flexible approach called SemRank. In this method, with

the help of a sliding bar user can easily vary their search mode from conventional search mode to discovery search mode.

Shahdad Shariatmadari et al.[14] present a method for finding semantic association based on the concept semantic similarity. ρ-operator [4] is used for discovering semantic similarities and graph similarity approach [15] is used to rank the similarity. The similarity between two paths will be calculated based on the degree of similarity of the nodes and edges using subsumption function proposed by Aleman-Meza [2]. The ranking approach proposed by Anyanwu,K. et.al [4] considers 'context' based on value assignments for different ontologies.

Aleman-Meza et al.[2] discuss a framework that uses ranking techniques to identify more interesting and more relevant semantic associations and define a ranking formula that considers Subsumption Weight S_P (how much meaning a semantic association conveys depending on the places of its components in the RDF), Path Length Weight L_P (that allows preference of either immediate or distant relationships), Popularity P_P (number of incoming and outgoing edges), Rarity R_P (rarely occurring entity) and Trust Weight T_P (determining how reliable a relationship is according to its origin) and context weight, for assessing the effectiveness of the ranking scheme. In this method '*user defined weight*' is assigned for each 'context regions' specified by the user and it is used to calculate the context weight. The ranking results depend on the criteria defined by the user.

Lee, M et al.[11] propose a semantic search methodology for measuring the information content of a semantic association that consists of resources and properties based on information theory and expanding the semantic network based on spreading activation. In this method, they provide search results that are connected and ordered relations between search keyword and other resources as link of relation on semantic network.

Dong, X., et al.[7] present a prototype system called Chem2Bio2RDF Dashboard for automatic collecting semantic association within the systems chemical biology space and apply a series of ranking metrics called Quality, Specificity and Distinctiveness to select the most relevant association.

To discover semantic associations between linked data, Vidal,M et al.[17] propose an authority-flow based ranking technique that is able to assign high scores to terms that correspond to potential discoveries, and to efficiently identify these highly scored terms. They also propose an approximate solution named graph-sampling. This technique samples events in a Bayesian network that models the topology of the data connections and it estimates ranking score that measure how important and relevant are the associations between two terms.

Major difference between our approach and other existing methods is that there is no facility provided to find the semantic association paths with user interested components are closer to either left or right entity. In our method, we have considered this feature in selecting choice for user's context which may be closer to either left entity or right entity. So, when paths are to be ranked, according to the users' choice, the discovery process finds semantic association paths with more context weighted entities are closer to either left or right entity is considered as highly relevant in ranking, while others are ranked lower.

4 Calculating Context Weight

Context Weight is one of the semantic metrics which is used to determine the relevancy based on a user specific view. Consider the scenario in which someone is interested in discovering how two persons are related to each other in the domain of "Funding Company". Concepts such as "Finance" or "Financial organization" would be most relevant, whereas something like "Music Company" would be less meaningful. So it is possible to capture a user's interest through a Context Specification through user interface screen. Thus, using the context specified, it is possible to rank a path according to its relevance with a user's domain of interest. Fig.1 illustrates various paths containing different domain entities in the RDF. The top most path (call it Path1) contains one Financial entity and one Music entity. The next path (call it Path 2) contains one Financial entity and other one not in the any domain category. The third path (call it Path 3) contain two Music entities. We assume that there are two users, user1 is more interested in Music domain and user2 is more interested in Financial domain. The expected ranking of these three paths for the user1 would be Path 3, Path 1, and Path 2 and for the user2 would be Path 2, Path1, and Path3.

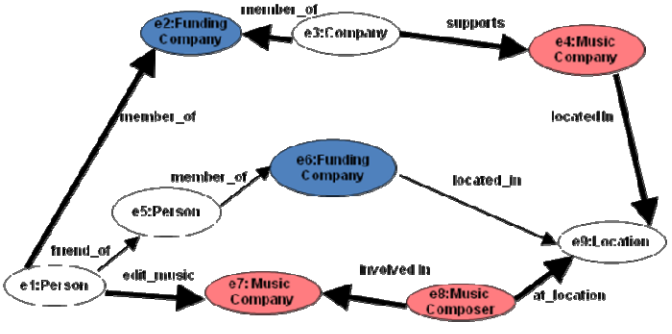


Fig. 1. Paths between two entities person and location in the RDF

From the Fig.1, paths may pass through more than one domain specified by the user. So the component (entities and properties) of the path passing through the region is multiplied by the corresponding weight of the region. Some of the components may not pass through any of the user specified region. These components may be irrelevant to the user. So we have to exclude the component for calculating context weight of the path. Given this background Aleman-Meza et.al[2] defined a formula to calculate the context weight of a given path P as follows:

$$C_p = \frac{1}{|c|}\left(\left(\sum_{i=1}^{region}\left(r_i\left(\sum c \in R_i\right)\right)\right) X\left(1-\frac{\#c \notin R}{|c|}\right)\right) \quad (1)$$

Here r_i is the user assigned weight of the region R_i and |c| is the total number of components in the path (excluding the start and end entities). To best of our knowledge Aleman-Meza et.al[2] proposed a method for ranking the semantic association using various criteria such as *Path length, Subsumption, Context, Popularity, Rarity and Trust* to get the relevancy.

Suppose the sub graph in an RDF contains paths between two entities and all the paths having same number of intermediate entities and are scattered in different position. In such case, Aleman-Meza et.al[2] method rank these paths arbitrarily, because of same context values. So, we have to rank this subset according to the users' interest. This can be achieved by using context closeness in ranking. We have defined the context closeness as follows:

Definition 5 (Context closeness): There exists a sequence $e_1,P_1,e_2,P_2,e_3,P_{3.....}$ e_{n-1}, P_{n-1},e_n in an RDF graph where e_i ($1 \leq i \leq n$) are entities and P_j ($1 \leq j < n$) are properties, and sum of context weight of user interested entities in the first half of the sequence is greater than second half of the sequence, we can say that the context is closer to left entity e_1; otherwise, context is closer to right entity e_n.

To calculate the *context value* based on the choice of the user selection in context closer, the formula (1) has been modified as follows:

$$C_t = cw_t \sum c \in D_t \tag{2}$$

$$MC_p = \frac{1}{|c|}\left(\sum_{t=1}^{|c|/2} C_t + 0.1 \sum_{t=|c|/2+1}^{n} C_t\right) X\left(1 - \frac{\#c \in D}{|c|}\right) \tag{3}$$

$$MC_p = \frac{1}{|c|}\left(0.1 \sum_{t=1}^{|c|/2} C_t + \sum_{t=|c|/2+1}^{n} C_t\right) X\left(1 - \frac{\#c \in D}{|c|}\right) \tag{4}$$

Where |c| is the total no of components in the path (excluding the start and end entities). The formula (3) and (4) are used to calculate the context weights closer to left entity and right entity respectively. The context weight MC_p is used as a parameter to calculate the weight of the semantic association paths.

4.1 Ranking the Semantic Association Paths

Our approach defines a path rank as a function of various intermediate weights. These are described as follows:

Subsumption weight **Sp:** In RDF, entities that are in the lower hierarchy can be considered to be more specialized instances of those further up in the hierarchy. Thus, lower entities have more specific meaning. So, high relevance can be assigned based on subsumption.

Path Length Weight L_P**:** In some queries, a user may be interested in the shortest paths. This may infer a strong relationship between two entities. In some cases a user may wish to find indirect or longer paths. Hence, the user can determine which association length influence.

Popularity Weight P_P**:** The number of incoming and outgoing relationships of entities called popular entities. Path contains highly popular entities may be more relevant. Hence, the user has to select either 'favor more popular associations' or 'favor less popular associations' based on their need.

Rarity Weight R_P**:** Sometimes rarely occurring events are considered to be more interesting than commonly occurring ones. Depending on the requirements, user has to select 'favor rare associations' or 'favor more common associations.[1][12].

Trust Weight T_P**:** Various entities and their relationships in a semantic association originate from different sources. Some of these sources may be more trusted than others. Thus, trust values need to be assigned to the meta-data extracted depending on its source.

We calculate Subsumption Weight S_P, Path Length Weight L_P, Popularity P_P, Rarity R_P and Trust Weight T_P [1] along with user-specific context weight MC_P. These weights are used to determine the path relevancy. So, all the intermediate weights are added to calculate the rank of each path.
Overall association Rank is calculated using the criteria as

$$W_p = k_1 \times S_p + k_2 \times L_p + k_3 \times MC_p + k_4 \times T_p + k_5 \times P_p + k_6 \times R_p \quad (5)$$

where $k_i (1 \leq i \leq 6)$ are preference weights and $\sum k_i = 1$. The resulting paths are ranked based on the users' domain of interest. Depending on the requirements, users can also change the preference weights to fine-tune the ranking criteria. In our experiments, we have given high weights to context component and use the other ranking components as secondary criteria.

5 Experimental Evaluation

For finding semantic association paths, we have used an RDF consisting of 52 classes, 70 properties and 3000 entities covering various domains such as Music, Finance, Terrorism and Sports etc. To test the performance of our system, we have selected 40 pairs of entities in the RDF. Semantic association paths has been generated and ranked under the various criteria such as favor short association or favor long association, favor popularity entities or favor unpopular entities, favor rarity, context closer to right entity or context closer to left entity. Criteria have been selected through user interface. Semantic association paths ranking has been done by the above users through the system as well as manually.

5.1 Preliminary Results

To demonstrate our ranking scheme's effectiveness, Fig. 2 shows comparison of human and proposed system ranking results between the entity sets (**Entity1:** John and **Entity2:** Slumdog millionaire). Here 'John' is entity under the class 'Music director' and 'Slumdog millionaire' is the entity under the class 'Movie'. The x-axis represents semantic associations rank first, second and so on according to the

proposed system results. The y-axis represents user-human ranking which is assigned manually by the users. We used the Spearman's footrule [6] distance as the measure of similarity between proposed system ranking and user-human ranking using the formula given below:

Spearman's Foot rule distance

$$D_{(system,human)} = \sum_{i=1}^{n} \left| R_{i_{system}} - R_{i_{human}} \right| \tag{6}$$

$$\text{Spearman's Foot rule Coefficient } C = 1 - \frac{4D}{n^2} \tag{7}$$

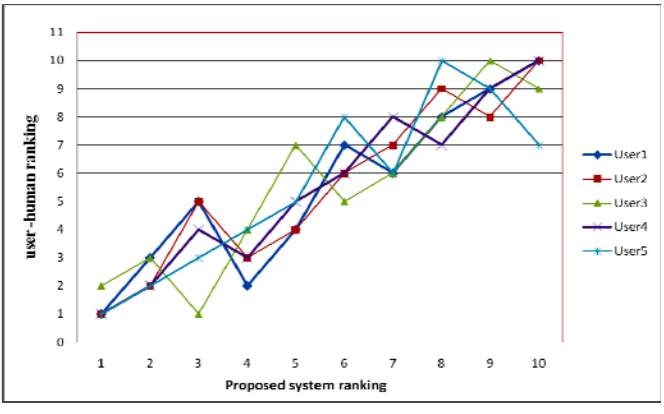


Fig. 2. Comparison of human and proposed system ranking with top-k results

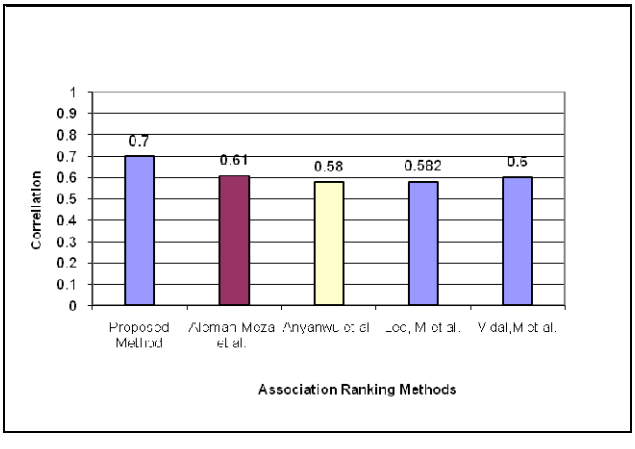


Fig. 3. Comparison of correlation

According to our experiments, the average correlation coefficients between proposed system ranking and user-human‘s ranking is 0.70. . Since the average correlation coefficient is greater than 0.5, the proposed system’s ranking and user-human ranking are highly correlated. Fig. 3 shows the comparison of Correlation between human rankings with proposed system and other existing association ranking methods. It explains that correlation between human ranking and our proposed approach is higher than other existing methods. We have evaluated the precision rate from the top-k semantic association paths from the ranked results for the proposed system and other existing methods. Precision represents the fraction of the relevant paths from top-k semantic association paths. Fig. 4 shows the comparison of precision rate of proposed method with existing methods. Irrespective of ‘k’ value, precision rate will increase or decrease. Among the five methods which show the same phenomenon. But,the method which we have adapted is more significant and provides high precision rate.

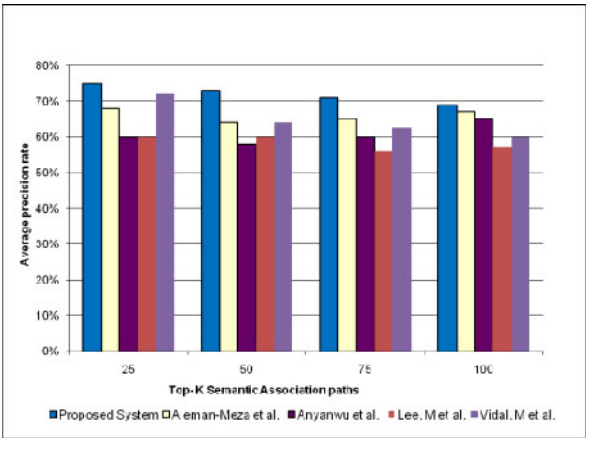


Fig. 4. Comparison of precision rate

6 Conclusion

Semantic data contains entities and heterogeneous relationships among them. The number of relationships between entities might be much greater than the number of entities. Ranking these relationship paths are required to find the relevant relationships between entities with respect to user’s domain of interest. Sometimes users’ may expect the relationships between two entities in which his/her context is closer to any one of the end points either left entity or right entity. The proposed method, find and rank the semantic association paths with respect to specific domain components which is closer to either left entity or right entity. We compare our proposed method with existing methods through spearman correlation coefficient and

precision rate. The average correlation coefficient between proposed system ranking, Aleman-Meza et al., Anyanwu et al., Lee,M. and Vidal,M., with human ranking are 0.70, 0.61,0.58,0.582 and 0.6 respectively. It explains that our proposed system ranking is highly correlated with human ranking. According to our experiments as measure of precision rate, we can conclude that our proposed system achieves high precision rate with top-k ranking than others. In future, we plan to generate the semantic web usage ontology from web usage information of each user and which may be used to get personalized semantic associations ranking.

References

1. Anderson, R., Khattak, A.: The Use of Information Retrieval Techniques for Intrusion Detection. In: Proc. 1st Int'l Workshop Recent Advances in Intrusion Detection (1998)
2. Aleman-Meza, B., Halaschek, C., Arpinar, I.B., Sheth, A.: Ranking Complex Relationships on the Semantic Web. IEEE Internet Computing 9(3), 37–44 (2005)
3. Anyanwu, K., Maduko, A., Sheth, A.: SemRank: Ranking Complex Relationship Search Results on the Semantic Web. In: Proc. of the 14th Int'l World Wide Web Conference, pp. 117–127. ACM Press (2005)
4. Anyanwu, K., Sheth, A.: ρ-Queries: Enabling Querying for Semantic Associations on the Semantic web. In: Proc. of the 12th Int'l World Wide Web Conference, pp. 690–699 (2003)
5. Berners-Lee, T., Hendler, J., Lassila, O.: The Semantic Web: a new form of web content that is meaningful to computers will unleash a revolution of new possibilities. Scientific American 285(5), 34–43 (2001)
6. Diaconis, P., Graham, R.: Spearman's Footrule as a Measure of Disarray. J. Royal Statistical Soc. Series B 39(2), 262–268 (1977)
7. Dong, X., Ding, Y., Wang, H., Chen, B., Wild, D.J.: Ranking semantic associations in systems chemical biology space. In: 19th Int'l World Wide Web Conference, FWCS 2010, Raleigh, NC (2010)
8. Jiang, X., Tan, A.: Learning and inferencing in user ontology for personalized Semantic Web search. Information Sciences 179, 2794–2808 (2009)
9. Lassila, O., Swick, R.R.: Resource Description Framework(RDF) Model and syntax specification, W3C Recommendation (1999)
10. Brickley, D., Guha, R.V.: Resource Description Framework (RDF) Schema Specification 1.0, W3C Candidate Recommendation (2000)
11. Lee, M., Kim, W.: Semantic Association Search and Rank Method based on Spreading Activation for the Semantic Web. In: IEEE Int'l Conference on Industrial Engineering and Engineering Management, pp. 1523–1527 (2009)
12. Lin, S., Chalupsky, H.: Unsupervised Link Discovery in Multi-Relational Data via Rarity Analysis. In: Proc. 3rd IEEE Int'l Conf. Data Mining, pp. 171–178. IEEE CS Press (2003)
13. Nazir, S., Afzal, M.T., Qadir, M.A., Maurer, H.A.: CAOWL-SAI: Context aware OWL based semantic association inference. In: NCM 2010, pp. 746–751 (2010)
14. Shariatmadari, S., Mamat, A., Ibrahim, H., Mustapha, N.: SwSim: Discovering semantic similarity association in semantic web. In: Proc. of the Int'l. Symposium on IT Sim 2008, pp.1–4 (2008)

15. Sokolsky, O., Kannan, S., Lee, I.: Simulation-Based Graph Similarity. In: Hermanns, H. (ed.) TACAS 2006. LNCS, vol. 3920, pp. 426–440. Springer, Heidelberg (2006)
16. Stojanovic, N., Mädche, A., Staab, S., Studer, R., Sure, Y.: SEAL – A Framework for Developing SEmantic PortALs. In: K-CAP 2001 – Proceedings of the First Int'l. ACM Conference on Knowledge Capture, Victoria, B.C., Canada (2001)
17. Vidal, M., Rashid, L., Ibabez, L., Rivera, J., Rodrogiez, H., Ruckhaus, E.: A Ranking-Based Approach to Discover Semantic Association Between Linked' Data. In: The 2nd Int'l Workshop on Inductive Reasoning and Machine Learning for the Semantic Web, pp. 18–29 (2010)

The Quantum of Language: A Metaphorical View of Mind Dimension

Svetlana Machova and Jana Kleckova

University of West Bohemia,
Department of Computer Science and Engineering, Univerzitni 8,
306 14, Plzen, Czech Republic
{smachova,kleckova}@kiv.zcu.cz

Abstract. In this article, we are focusing on metaphor of mental states examining hidden dimension of the human mind and its interaction with a world. Our natural language research methodology were included constructing a database of metaphorical stimuli, randomly presenting these to participants, and tabulating FAE responses both by participant and by the stimulus, analyzed by frequency of semantic type. The project aimed to produce a new model of knowledge representation. Acquired data suggest that mind's metaphoric self-reflection semantics closely correlate with entangled quantum particles and light binding phenomena. Created semantic network gave us a unique opportunity to visualize the probabilistic picture of mind itself within an interdisciplinary approach of cognitive and computational networking studies.

Keywords: Semantics, NLP, Free Associations, Quantum linguistics, Mind.

1 Introduction

This paper proposed a new method representing a cognitive dimension of mind based on the psychoanalytic methodology of Free Association Test and Lakoff's cognitive metaphor theories. Common sense view of mind self-perception is an important tool representing a general approach specifying default reasoning.

There are several models describing how the mind works in the realm of ideas. The first is a W.Humboldt's theory describing human speech ability as a spirit of language. The second is an A. Nalimov's probabilistic idea of the semantic fields. The next is implemented reasoning model called ATT Meta "performs a type of metaphor based reasoning" developed at Birmingham University. [2, 3, 12]. J. Barnden is arguing that metaphor is a particularly significant field in mind studying as its most of the times the only way how a mind can describe itself. [2, 3, 4].

According to J. Barnden, fully/fledged science can describe a mind as it describes itself in natural language [7, 12, 13]. English based metaphoric concepts such as MIND PARTS AS PERSONS eg.*"One part of Mike knows that Sally has left for good"* or MIND AS PHYSICAL SPACES eg. *"inner/outer Self", "Emptiness was killing herL,* were as well present in Czech data. Eg. *"Citil se rozpolceny, roztrzity", "Jeho pohled byl prazdny, okna do duse nevyjadrovali zhola nic, jako by odesel."*.

J.-S. Pan, S.-M. Chen, N.T. Nguyen (Eds.): ACIIDS 2012, Part III, LNAI 7198, pp. 163–168, 2012.

In our latter researches [10, 11] was shown that metaphors of mind significantly correlate with Cartesian view of mind / body duality. Both theories, claiming unity duality of mind and body nature, can be true (see out of body metaphor Fig.1). It is two sides of a problem, - two major states of mind: consciousness presence – transcendent unity and MIND WANDERING (non-presence) duality. [13].

The Quantum Mind Project (QMP) is knowledge representation research that uses Free Associations Test data to create a semantic network of Mind from associations to a unique list of cognitive metaphor stimuli. QMP is a fundamentally new resource of deep semantic knowledge, a platform for mind study and optimizing the human communication process with electronic resource.

According to K. Ahrens [1] there is a need for cognitive models reflecting the relationship of the Mind within and out of the physical world. As W. James noticed we are divided into "spiritual me, social me and material me"[1, 12]. Conceptual Metaphor Theory rejects the notion that metaphor is a decorative device and suggest that metaphor structure thinking and knowledge [7, 8, 9].

2 Methods

In contrast with already existing WordNets or MindNET Microsoft research [17] based on collocations or existing corpuses processing, these algorithms work on immediate authentic respondent answers for given text stimuli. To create and visualize association semantic network, we used a set of web application tools. A database was generated in SQL by an original algorithm in on-line mode. Semantic dependency data were loaded from the CSV file and visualized with Gephi dynamic graph. We also examined which stimuli-response pairs occurred more frequently, which responses occurred more often, were there any unexpected connections between certain words? [12]

A list of 400 stimuli for Free association Test in the Czech language has been created. In the database, we have collected 12,673 word responses from 200 participants. Most of the participants were university students, mostly in humanities field such as history or philosophy. The stimuli appeared to each respondent in a random order, and she/he had the opportunity to write 0-3 responses (association) to each stimulus. Respondents had to write the first word or phrase which immediately came into their mind. [10, 11, 12].

3 Experimental Findings

Stimuli -association strings from the database were sorted into several main semantic spaces according to associations:

1. *emotions* (negative/ positive, love, excitement, friendship, feelings, emptiness). According to Picard "people have skills for detecting when someone is annoyed or frustrated and for adapting to such affective cues. Three factors are especially important: perceiving the situation, perceiving affective expression, and knowing how interrupting at such a time was received in the past. The ultimate

strategy involves more than affect perception, but affect perception is critical." [14]. The conceptual metaphor for emotion - HAPPINESS IS LIGHT, writes Kovecses [1, 12] such as "To shine with happiness", "His face became dark when he realized that"-, "Let shining your truth colors," "Her personality is shining through" supports the idea of inner Self as something, which is "enlightened" by personality (eg. *bright* Mind) and colored by emotions.

2. *mind's puzzling* (distraction, dispersion/concentration, torn into pieces, mind behaving as an aeter or dissolving gas. (see examples and Fig. 1 below)

3. *mind's in- and out of body* experience (being out of self, to be spaced out, to be taken, being not myself (Divided person metaphor [3]), feeling strange, alienation, flying, being beside myself with smth. The other important concept is a MIND AS A SPACE [2, 4] or MIND WANDERING metaphor, such as "I was beside myself with anger", I am not myself today" [8], "I have got you under my skin" were occurred in a participants responses the most frequently.

4. *mind's interaction* (empathy, interest, openness, clearness). It is obviously not only in English that cognitive realm representing its states mentioning: "fringes of consciousness," about "ideas surfacing at the mind", that we can see mental things "clearly" or "obscurely" [1].

Mind -ontology of interaction.

Table 1. Metaphoric examples of Mind as a Quantum phenomen

Czech	Russian	English	Literal English translation
Se-tkání	Vstreca	Meeting	Co-spinning, co – binding
Pařit	Zavisat	Get high,	Hanging in the air
Bavit se	Baldet, rastvoritsja v tolpe	To talk, having fun	Mixing, dissolve in a crowd
Roz-loučit se	Razluka	Say good buy	Divide light rays
Vy-točit nekoho (slang)	Zakolebat kogo-to (slang)	Get angry	To spin out (of the body), make someone waving
Souznit, byt naladen	Garmonija	Harmony	Stay tuned
Vytvorit pouto, zaplest se	Svjazatsja,	Attachment	To create bond, tie
Vytvorit spojeni	Prisojedinitsja	Connection	Get involved , mixed up

Unique phenomena are described by metaphor of mind's communication networks: it emphasizes *unity*, *creating of atmosphere*, network, and how the mind *in-flu-ence* another behaving as an enlighten fluid/stream of consciousness. According to mind's self-reflection it is also able to "*mix"* (with each other), to *build bonds, to interfere waves* [6] as it were explained in [12, 15]*, and "attach" strings to each other.* All those phenomena are universal in all languages we have studied so far.

As we mentioned in previous studies [10, 11], the concept of *influence* on a deep semantic level suggests a functional similarity with *inspiration*. Moreover, the in-*flu*-ence root "flu", in Russian *vlijanie*, in Czech *.vliv*, metaphorically depicts a certain kind of consciousness stream *"in-spiring"* influenced person. The research seems to confirm the Cartesian dualistic theory about mind possessing, leaving or transcending the body. These associations seem to represent conceptual metaphors of Lakoff's CONTAINER conceptual metaphor [7] *being out of body, inner self as a light, fluid, gas, life as a spinning*.

We have classified our data from different points of view; we were after all interested in the qualitative findings. Unique associations produce another cognitive metaphor in connection with these stimuli. Primarily, we were interested in creating a semantic network from obtained data. The example of the network describing a concentrated/distracted state of mind we can see from the Fig 1 below. Emphasized arrows mean s semantic dependency. [See also 10, 11,12]

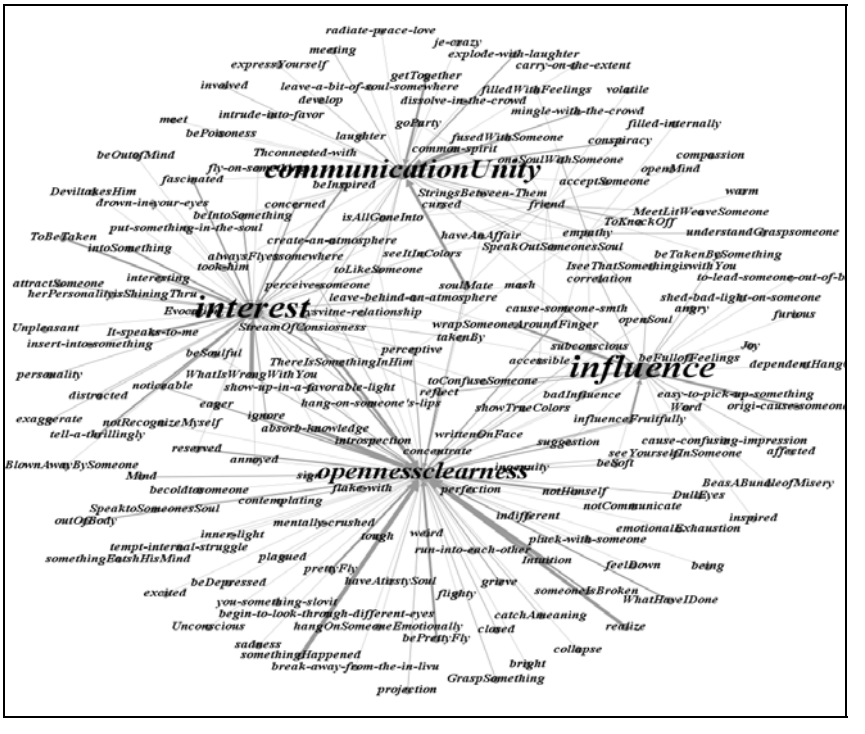


Fig. 1. Ontology of Mind,s interaction. Semantic network.

As we can see from the Tab. 1.metaphorical view suggests that our Self (Mind) somehow, on a quantum level, physically *entangled into* a communication network with others. See also [15, 16]. The *communication unity, interest, influence, openness and "clearness" of mind* are the main concepts providing an ontological view of Mind's interaction and to create *common* spirit or *atmosphere* and to be able to communicate we literally need to open our Mind. Eg. *"I always accepted (let in) him as a person."*

4 Conclusion

Empirical findings about the phenomenology of mind gained from Associative Experiment, has opened a new dimension of artificial intelligence and quantum linguistics research, allowing us to create a link between a collective unconscious level and a conscious Mind. In this study we attempted to explore and visualize the mind and its interaction. Examining of metaphorical reasoning concepts is extremely important not only for creating algorithms implemented in artificial intelligent agents, but also to explain well-known psychoanalysis and psychiatry phenomena, such distraction, multiple personality states of mind or some neurogenic speech disorders.

Understanding of the metaphoric self-reflection might be guided by understanding how the quantum like entanglement influence socializes and structures our natures, how it shapes our mental life's architectures in Czech and other languages. We realize that every language is a complementary puzzle of this latent semantic information, which can be possible discovered. As a validation of our research we aim to extend this experiment on English and Russian languages and compare the metaphoric view of mental states with original data. As research suggests, literal meaning of metaphor is a huge field of research discovering hidden structures of the mind helping computational intelligence to achieve a natural human - like reasoning.

Acknowledgments. The Elegant Mind project presented in this paper is supported by the Czech Science Foundation grant GACR 106/09/0740.

References

1. Ahrens, K.: Embodying the Self: A Linguistic Analysis. In: Huang, S.F. (ed.) Proceedings of the International Symposium on Body & Cognition: A Multidisciplinary Perspective, pp. 22(1)–22(19). National Taiwan University, Taipei (2005)
2. Barnden, J.: ATT-Meta Project Databank: Examples of Usage of Metaphors of Mind (2010), http://www.cs.bham.ac.uk/~jab/ATT-Meta/Databank/Egs/Text/Literature/saturday.ian-mcewan
3. Barnden, J.: Consciousness and Common-Sense Metaphors of Mind. In: Nualláin, S., et al. (eds.) Two sciences of Mind: Readings in Cognitive Science and Consciousness. John Benjamins Publishing Company (1997)
4. Barnden, J., Helmreich, S., Iverson, E., Stein, G.: Artificial Intelligence and Metaphors of Mind: Within-Vehicle Reasoning and its Benefits (1996)
5. Deignan, A.: Metaphor and Corpus Linguistics. University of Leeds (2005)
6. Fournier, J.M.: Light binding, Rowland Institute, Harvard (2003), http://www.rowland.harvard.edu/organization/past_research/optics/default.html
7. Lakoff, G., Johnson, M.: Metaphor We Live By. University of Chicago Press (1980)
8. Lakoff, G.: Sorry, I'm not Myself Today: The Metaphor System for Conceptualizing the Self. In: Fauconnier, G., Sweetser, E. (eds.) Spaces, Worlds, and Grammar. University of Chicago Press (1996)

9. Lakoff, G., Johnson, M.: Philosophy in the Flesh. The Embodied Mind and Its Challenge to Western Thought. Perseus Books, N.Y (1999)
10. Machová, S., Kratochvil, P., Klečková, J.: The Elegant Mind. A New Insight To The Deep Semantic Network. In: Annals of DAAAM for 2010 & Proceedings of the 21st International DAAAM Symposium "Intelligent Manufacturing & Automation: Focus on Interdisciplinary Solutions", Zadar (2010)
11. Machova, S., Kleckova, J.: Are we Waves or are we Particles? A New Insight into Deep Semantics in Natural Language Processing. In: IEEE NLP KE 2010, Proceeedings of the 6th International Conference on Natural Language Processing and Knowledge Engineering, Bejing (2010)
12. Machova, S., Kleckova, J.: The Quantum of Language. Natural semantics as a Mirror of Consciousness. In: Proceedings of the International Conference on Pervasive, Embedded Computing and Communication, ICPECC 2011, Hong Kong (2011) (in print)
13. Machova, S., Kleckova, J.: The Quantum of NLP. Cognitive Metaphor as a Mind Discovering Device. In: International Joint Conference on Biomedical Engineering Systems and Technologies, HEALTHINF 2012, Alhavre (2012) (in submission)
14. Picard, R.W.: Affective Computing: Challenges. International Journal of Human-Computer Studies 59(1-2), 55–64 (2003)
15. Radin, D.I.: Entangled minds: extrasensory experiences in a quantum reality. Simon and Schuster (2006)
16. Sheldrake, R.: The sense of being stared at: and other aspects of the extended mind. Arrow (2004)
17. Vanderwende, L., et al.: MindNet: an automatically-created lexical resource. In: HLT/EMNLP Interactive Demonstrations Proceedings (2005)

An Experimental Study to Explore Usability Problems of Interactive Voice Response Systems

Hee-Cheol Kim

Dept of Computer Engineering/UHRC, Inje University, Obang-dong 607,
Gimhae, Gyeong-Nam, 621-749, S. Korea
heeki@inje.ac.kr

Abstract. While interactive voice response (IVR) systems employing touch tone interface (TTI) are popularly used these days, they are generally known for their inconvenience. This is not only because of the characteristics that TTI inherently has, but also because of lack of understanding of IVR system users. This study is aimed at contributing to capture an understanding of the users, which eventually leads to better system design. In particular, we have developed an IVR system simulator to enable efficient, flexible, and rich usability tests for IVR systems. This paper presents an experimental study on usability of IVR systems utilizing the simulator. In the experiment, 41 subjects performed three different tasks concerning phone charges using the simulator to identify various usability problems. The analytic results are hopefully a basis for user-centered IVR systems design.

Keywords: Human Computer Interaction (HCI), Interactive Voice Response System (IVR), Simulation, Touch Tone Interface (TTI), Usability.

1 Introduction

Computer telecommunication integration (CTI) technology includes such applications as auto-answer, voice mail, fax, short-message service (SMS), automatic switching, sound recording services, etc. Nowadays it is an indispensable technology both in our everyday lives and in the competitive environment of modern business.

IVR systems are an important part of CTI. Users can access required voice information using a touch-tone interface (TTI; see Fig. 1). There are a large number of IVR systems-based applications with this simple interface, e.g. to check orders and reservation status, carry out surveys, and pay for and transfer fees, etc.

The body of literature about the usability of IVR systems provides numerous solutions for overcoming the various deficiencies of IVR systems. For instance, Gardner-Bonneau suggested balancing the depth and width of a menu by designing a graph structure [1]. One found that including an additional component increases the controllability of IVR systems [2], and skipping and scanning keys can be used to reduce navigation time [3]. Some systems allow users to set personal menus and to access them directly in the hierarchy without going through the various layers [4][5]. Insertion of speech-based earcons or non-speech earcons reportedly improves

J.-S. Pan, S.-M. Chen, N.T. Nguyen (Eds.): ACIIDS 2012, Part III, LNAI 7198, pp. 169–177, 2012.

navigation through auditory menus [6][7][8]. While these approaches can help increase convenience of IVR systems to some extent, however, they are technique-oriented solutions. To date, few usability-related studies have been conducted, and the level of user satisfaction with IVR systems is still low. In fact, many usability problems in IVR systems have been reported. For example, users are likely to get lost when the paths are ambiguous [9][10], and have a great difficulty to sense an overview of the menu structure. In many cases, they do not find the required information even using the entire menu and all system prompts. Further, interaction with IVR systems has some inherent usability problems such as linearity, transcience, ambiguity, and minimal feedback [11].

Fig. 1. Touch-tone interface

This paper reports an experimental study to explore usability problems of IVR systems utilizing an IVR simulator, where all 41 subjects participated by performing three different tasks. Here, the simulator was designed to help efficient and rich usability tests for IVR systems. The simulator is very useful, because it can be used to build an experimental user test in a more flexible way, and provides the statistics and data about users' behaviors recorded during the test.

Section 2 describes the objective and the method of this research including the experimental setting. Section 3 presents the IVR system simulator that we have developed in terms of why it is needed, what it is, and how it is used. The results of the user test are finally presented in section 4, followed by a conclusive section.

2 Objective and Method

Until now, IVR systems design has been technique-oriented. This orientation has a certain limitation to resolve inconvenience of their usage. In that IVR systems are commonly used, however, it is time to understand callers or users more than techniques. In relation to this, we present the objective and the method for this study.

2.1 The Objective

The purpose of the study is to understand IVR system users. More specifically, we want to explore various usability problems in them, which will be a basis for usable IVR system design and better and deeper understanding of IVR system users in the future.

In particular, the IVR system simulator was designed to use for the special purpose of usability tests. The simulator overcomes traditional user studies adopting real IVR systems where it is impossible to test various contexts and tasks, because one cannot change the menu structure and terms of the real IVR systems. It means that the simulator provides flexible ways of studying users by easily defining and modifying tasks that callers carry out and contexts in which they are.

2.2 The Method

We have established an experimental setting in which 41 subjects participated to carry out three different tasks using the simulator. They were all undergraduate students recruited from the department of computer engineering, Inje university, Korea. Three tasks concerning call charge were given: "Get the information about the amount of free calls and SMS", "Get the information about types of payment systems for call plan", and "Get the information about the telephone bill that you will pay for this month."

Each subject was asked to input his or her personal information, and perform three tasks in succession using the simulator. Their actions, particularly about the path to which they moved, and the elapsed time were recorded. After completing the tasks, they also filled the forms of the questionnaire in relation to user satisfaction, ease of understanding terms, sense of the menu structure, and so on. The actions, the path, the elapsed time, and the answers to the questionnaire logged in the simulator were later analyzed.

As a final step, we had small interviews with them, asking them to explain each movement, telling why they chose the path and what usability problems are. Here experimenters made note of the interview contents and they were analyzed.

3 IVR System Simulator

As a system for IVR usability researchers, The IVR system simulator provides an environment for the design of a mock task and users' performance of the task. Further, it saves the history of a mock task and analyzes users' action patterns to some extent. By using the simulator, we expect that researchers and practitioners will gain some help to understand IVR system usability and design the IVR system interface. In this section, we present the simulator's structure and functions.

3.1 Simulator Structure

The IVR system simulator consists of (1) an experimenter's application to design experiments, (2) a callers' (or subjects') application to perform the task, (3) and a data analysis application. Firstly, the experimenter's application is used to design the workflow of a mock task, and design a questionnaire. Secondly, the caller's application is the one enabling to perform a task, and to record in the DB callers' actions, the elapsed time, the path that they take, and answers to the questionnaire. Finally, the data analysis application provides callers' actions and the history of task performance recorded during the experiment, as well as basic statistics. Figure 2 shows an example of UI of the caller's application in the IVR system simulator. The simulator was developed with Java, as well as eclipse, MySQL, Navicat, PowerDesigner, and IIS.

Fig. 2. An example of caller's application interface in the IVR system simulator

3.2 Functions of the Simulator

The simulator can be best understood by functions it supports. We briefly present its functions according to three different applications.

Experimenter's application. There are three major functions in it. Firstly, it supports workflow design, by which the experimenter can define a task project name, and explain each node for the item. It also helps to define and organize the menu structure and voice contents. Secondly, the application helps to the task path by defining the last node of the given task. Thirdly, the experimenter can organize questionnaire forms, because the application enables to input questions and create different types of questions, e.g. Likert scale form, radio button form, open questions, and the like.

Caller's application. By this application, the caller inputs his or her general and personal information, which will be used during the data analysis phase. Using the interface shown in Figure 2, the caller performs a given task, and the information about the actions, the path, and related time are logged. Subjects also answer the questionnaire with the application.

Data analysis application. This application supports four types of analysis. Firstly, an analysis of callers is provided, by which one can confirm and analyze callers' personal information, and get it with a form of Excel. Secondly, there is also an analysis for workflow. There, the experimenter can see the callers' path, and sense an overview of the menu structure. Thirdly, there is also a task analysis module. About each task, one can observe the number of callers' mistakes of pressing wrong keys, the elapsed time for the task, and see callers' personal information. Fourthly, the experimenter can confirm basic statistical results on answers to the questionnaire, which can also be exported by an Excel format.

4 Results and Analysis

There are two parts in the analysis: Firstly, a general analysis was conducted in terms of some important usability criteria that Lewis proposed [12] such as task completion time, task completion rate, caller satisfaction, sense of menu structure, and the like. Secondly, we analyzed the interview data to identify usability problems. The following are the three given tasks.

- Task 1: Get the information about the amount of free calls and SMS.
- Task 2: Get the information about types of payment systems for call plan.
- Task 3: Get the information about the telephone bill for this month.

4.1 General Analysis

The general analysis in the study has been done on the basis of the statistics and caller answers to the questionnaire recorded in the simulator.

Task completion rate and time. All subjects completed the tasks without failure, partly because the given tasks were neither very complex nor ambiguous. However, callers did not always find optimal paths to success (optimal completion). Rather, they made some mistakes, repaired them, and finished the tasks (delayed completion). The table 1 shows the summary of the results about task completion rates with respect to optimal and delayed completion, and finally the elapsed time to complete tasks on average.

Table 1. A summary of task completion rate and time

Task	Optimal completion	Delayed completion	Completion time(sec) Optimal/Delayed
Task 1	30	11	26 / 56
Task 2	22	19	31 / 53
Task 3	33	8	25 / 39

Caller satisfaction. When subjects were asked about if they were satisfied with IVR systems usage, majority of them (19 callers) neither agreed nor disagreed with it. However, they tend to think that talking to call center operators directly is better than using IVR systems. Table 2 shows related results.

Table 2. A summary of caller satisfaction

Statement	Strongly Disagree	Disagree	Neutral	Agree	Strongly agree
I am satisfied with IVR systems usage	2	10	19	9	3
It is more convenient using IVR systems than talking operators directly	11	16	6	9	1

Sense of overview. Speech and audio interfaces are inherently sequential, whereas visual interfaces can be simultaneous [13]. It means that speech interfaces have a difficulty to sense an overview of the overall information space, while visual interfaces enable users to do it. In this respect, we had a hypothesis that callers feel inconvenience when sensing the menu structure with IVR systems, and so are likely to get lost in the information space expressed by the menu structure. However, the result shows that we need not take this hypothesis seriously (see Table 3). In fact, subjects thought that the menu structure was well-organized, and was neutral concerning the problem of getting lost in the information space.

Table 3. Sense of the menu structure and caller location in the information space

Statement	Strongly Disagree	Disagree	Neutral	Agree	Strongly agree
The menu structure is well-organized	6	5	16	16	3
I was aware of where I am in the menu hierarchy during the experiment	3	13	12	14	1

Perceived ease of understanding of terms. According to the results, terms and expressions used in the tasks were neither easy nor difficult to understand. As far as three tasks are concerned, this neutrality can imply that IVR systems are acceptable to use, but still unsatisfactory to some extent. This interpretation is supported by the results of the short interviews with subjects, which will be soon discussed.

Table 4. Degree of understanding terms and expressions

Statement	Strongly Disagree	Disagree	Neutral	Agree	Strongly agree
Terms and expressions used were easy to understand	4	11	14	13	1

4.2 Identifying Usability Problems

The simulator logs the elapsed time for each task. It also records the path that they take during the experiment, e.g. they press button 1, and then button 3. After tasks are performed, we particularly had short interviews with subjects along the path recorded, asking them to answer why they have moved by the path that they chose, what usability problems exist, and the like. By doing so, we wanted to understand usability problems of IVR systems. Here are typical usability problems that we identified.

Term ambiguity. Many subjects reported that ambiguous terms and expressions are reasons for delayed completion. All twelve subjects pointed it out (Three subjects for Task 1, five for Task 2, and four for Task 3). Further, there were nine cases where the

meaning of a term that subjects assume is different from its meaning that the system assumes (four for Task 1, three for Task 2, and two for Task 3). This mismatch between the user's and the system developer's perception concerning the same term is definitely a source of usability problems. This phenomenon is related to the one generally known as "gulfs of evaluation/execution" [14], or "ontological drift" [15] in HCI community. Also, nine subjects reported that some words were difficult to understand, i.e. they were not quite sure of what the terms mean (two subjects from Task 1, four from Task 2, and three from Task 3). In all, term ambiguity is a major usability problem that frequently occurred.

Phonetic deficiency. While analyzing the results, we found that three important dimensions should be considered about phonetic deficiency: pronunciation, volume, and voice speed. These three features are equally important based on the results. As we see the resulting analysis in Table 4, five cases about voice imprecision, 15 about weak volume, and 11 about fast voice speed were reported, as far as phonetic deficiency is concerned. Before the experiment, we did not expect phonetic deficiency as a primary usability problem. However, it turned out to be extremely important to be aware that users are very sensitive to pronunciation, volume, and voice speed.

Table 5. Usability problems concerning phonetic deficiency

Phonetic deficiency	Task 1	Task 2	Task 3	Total
Pronunciation imprecision	1	1	3	5
Weak volume	5	6	4	15
Fast voice speed	2	6	3	11

Information navigation. According to the data, there were many cases when callers have to go back to the root or the previous menu for one reason or another. Eleven subjects pointed out the problem of navigating the menu repeatedly (5 subjects from Task 1, 3 from Task 2, and 3 from Task 3), which is a main reason for delayed completion or for failure. Based on the data, the navigation problem over the menu structure is serious enough to take as a major usability problem. In fact, callers often face that there is no other way but going back to the previous state. Comparing with GUI, voice user interface (VUI) is inherently linear [16]. Here, when an interaction type is linear, it means that users navigate over information only according to a given order. For example, if a person listens to music by cassette tapes, (s)he cannot control tapes freely and easily when (s)he moves or navigates from the first music to the last in the tape, and so it takes much time to do it. The ear cannot browse around a set of recordings, unlike the eye, which can scan a screen of text and images at a glance. In this sense, the linearity problem is almost the same as the navigation problem in IVR systems. Therefore, it is important to consider the navigation (or linearity) problem when designing systems based on speech and audio interfaces.

Cognitive overload. From the data collected, pure user mistakes were also found, which are usually related to either cognitive or motor activity. For instance, seven

cases of keying errors took places (3 cases from Task 1, 3 from Task 2, and 1 from Task 1). Three subjects also reported that they have lost their concentrations, though each task took less than a minute only, which is very short (2 subjects from Task 2 and 1 from Task 3).

5 Conclusion

The wide acceptance of IVR systems is due to the industrial need to reduce costs by employing IVR systems for customer care on one hand, and to technological development on the other hand. In spite of their success, however, users' desire for convenient interaction with the systems has been largely ignored. As a matter of fact, few studies have focused on the usability or user-oriented design of IVR systems. Against the problem, this study has been conducted to understand users' needs and usability problems through an experimental user test. In particular, we have designed an IVR system simulator to enable efficient, flexible, and rich usability tests for IVR systems, and have used it for the experimental study.

Among many findings, term ambiguity was tuned out to be the most serious usability problem. Certainly, terms or expressions taken should be unambiguous and be crucial objects to consider in the IVR system design process. Phonetic deficiency concerning pronunciation, volume, and voice speed was also found to be a major problem, which was larger than we thought. Further, IVR systems taking VUI and TTI have an inherent problem called linearity or navigation problem. This problem was also confirmed by the data obtained. In summary, users were neutral towards usability or user-friendliness of IVR systems. We believe that this implies users are in a conflicting situation where while they use IVR systems in everyday lives, they are not yet satisfied with them. In this respect, more and much research on usability of IVR systems still remains. Finally, these findings can hopefully be used to help designers design more usable IVR systems.

Acknowledgments. This work was supported by the Korean Research Foundation (KRF) grant funded by the Korea government (MEST) (No: 2009-0076772).

References

1. Gardner-Bonneau, D. (ed.): Human factors and voice interactive systems. Kluwer Academic, Norwell (1999)
2. Karat, C.M., Halverson, M., Horn, D., Karat, J.: Patterns of entry and correction in large vocabulary continuous speech recognition systems. In: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, pp. 568–575. ACM Press, New York (1999)
3. Resnick, P., Virzi, R.A.: Skip and scan: Cleaning up telephone interfaces. In: Proceedings of ACM CHI 1992, pp. 419–426 (1992)
4. Shajahan, P.M., Irani, P.P.: Improving Navigation in Touch-Tone Interfaces. In: International Symposium on Human Factors in Telecommunication, Berlin, DE (2003) 20031201-20031204
5. Irani, P., Shajahan, P., Kemke, C.: VoiceMarks: restructuring hierarchical voice menus for improving navigation. Int. J. Speech Technol. 9, 75–94 (2006)

6. Walker, B.N., Nance, A., Lindsay, J.: Spearcons: speech-based earcons improve navigation performance in auditory menus. In: Proceedings of the 12th International Conference on Auditory Display (ICAD 2006), pp. 63–68 (2006)
7. Brewster, S.A.: Navigating telephone-based interfaces with earcons. In: 12th British Computer Society HCI Conference, Bristol, England (1997)
8. Shajahan, P., Irani, P.: Representing Hierarchies Using Multiple Synthetic Voices, pp. 885–891. IEEE Computer Society (2004)
9. Balentine, B.: Re-Engineering the speech menu. In: Gardner-Bonneau, D. (ed.) Human Factors and Voice Interactive Systems, pp. 205–235. Kluwer Academic, Norwell (1999)
10. Balentine, B., Morgan, D.P.: How to build a speech recognition application. Enterprise Integration Group Inc., San Ramon (1999)
11. Kim, H.-C., Liu, D., Kim, H.-W.: Inherent Usability Problems in Interactive Voice Response Systems. In: Jacko, J.A. (ed.) HCI International 2011, Part IV. LNCS, vol. 6764, pp. 476–483. Springer, Heidelberg (2011)
12. Lewis, J.R.: Speech user interface development methodology. In: Practical Speech User Interface Design, ch. 6, pp. 93–168. CRC Press (2011)
13. Arons, B.: Hyperspeech: Navigating in Speech-Only Hypermedia. In: HYPERTEXT 1991: Proceedings of the Third Annual ACM Conference on Hypertext, pp. 133–146 (1991)
14. Norman, D.A.: Cognitive engineering. In: Norman, D.A., Draper, S.D. (eds.) User Centered System Design, pp. 31–61. Lawrence Erlbaum Associates, Hillsdale (1986)
15. Robinson, M., Bannon, R.: Questioning representations. In: Proceedings of ECSCW 1991, pp. 219–223 (1991)
16. Kim, H.: Weaknesses of voice interaction. In: The 4th International Conference on Networked Computing and Advanced Information Management (NCM 2008), vol. 2, pp. 740–745 (2008)

Using Eyetracking in a Mobile Applications Usability Testing

Piotr Chynał, Jerzy M. Szymański, and Janusz Sobecki

Institute of Informatics, Wroclaw University of Technology
Wyb.Wyspianskiego 27, 50-370 Wroclaw, Poland
{Piotr.Chynal,Jerzy.Szymanski,Janusz.Sobecki}@pwr.wroc.pl

Abstract. In this paper we present general problems of a mobile application usability testing by means of eyetracking. The motivation for considering this problem is the fact that eyetracking is still one of the most advanced usability testing tool. We achieved that by performing two eyetracking tests with the participation of users. We tested mobile application on smartphone and PC emulator, to find out which method gives the most valuable results. Both tests showed that eyetracking testing of mobile applications gives valuable results but to make it really efficient professional equipment designed for mobile eyetracking is required.

Keywords: Eyetracking, Usability, Human-Computer Interaction.

1 Introduction

Modern information systems often suffer from many usability problems. Applications developed for mobile phones are no exception [11], [12]. The ISO9241-11 norm defines usability as "extent to which a product can be used by specified users to achieve specified goals with effectiveness, efficiency and satisfaction in a specified context of use" [8]. There are several well known techniques for the usability verification (for example focus groups, interviews, observations, surveys, etc.). One of the most interesting usability testing techniques is eyetracking [4], [9]. This method enables to track the movement of user gaze on the screen, using a special infrared camera called eyetracker. In result of such test we receive graphical reports of where users were looking during performing tasks in the application. This provides data for effectiveness and efficiency analysis. It has few disadvantages, such as motionless head during eye tracking, using a variety of invasive devices, a relatively high price of commercially available eye-trackers and a difficult calibration [3], [9]. However, it provides very valuable information for usability studies. All of them are based on the eye-mind hypothesis that what a person is looking at, is assumed to indicate the thought on top of the stack of cognitive processes.

The main purpose of our study was to verify the known eyetracking method in the considerably new, mobile environment. ComScore study [7] shows, that Smartphone adoption in the U.K., France, Germany, Spain and Italy has grown 41 percent in the past year (2010). The importance of mobile phones is increasing in the everyday life,

J.-S. Pan, S.-M. Chen, N.T. Nguyen (Eds.): ACIIDS 2012, Part III, LNAI 7198, pp. 178–186, 2012.

and the usability of mobile applications is becoming a critical factor. Nielsen Norman Group mobile applications usability studies conducted in 2009 showed that the average success rate on required tasks was only 59% [11], which is a very low percentage in terms of proper functioning of the application.

We decided to undertake research on the real mobile phone, with users using an existing application. The chosen mobile application should enable testing on both types of users, those who are familiar with it or its web equivalent, and those who are completely new to mobile applications. We have selected the mobile, touch screen version of facebook.com[1] web service, which is one of the most popular mobile applications, but while using it we found several usability issues in it.

2 Tools Used in the Experiment

Experiments were conducted in the Software Quality Laboratory, at the Wroclaw University of Technology. The main equipment used during this research was ASL 6000 eyetracking module [1]. It consists of two computers, Head Mounted Optics module (HMO) with camera, and ASL module with monitor. Study was performed using a chosen smartphone, a computer emulator and second camera for recording the phone screen. The brief description of the equipment and software is presented below:

1. ASL Eye-Trac 6000 Head Mounted Optics. The head mounted eye tracker accurately measures a person eye line of gaze with respect to their head. It is attached to the head by mounting on a headband (Fig. 1). The headband is light and adjustable to user head size [2]. One of the eyetracking cameras is using infrared to detect pupil and cornea reflections. The control unit processes the eye camera signal to extract the pupil and reflection data and computes both pupil diameter and line of gaze [1], [2].

Fig. 1. Testing environment

[1] http://www.touch.facebook.com

2. Web Cam Logitech Quick Cam Pro 9000. This camera was used to record the screen of the smartphone and operations made by users during interaction with the application. It was mounted as an additional element of HMO headband to assure closest angel to user's field of vision (Fig. 1). We used this camera, because parameters of the standard camera mounted in HMO were too low, to enable recording of the smartphone screen in the descent quality.
3. YouWave Android. This application was created by YouWave LLC. It is one of the easiest to use and most advanced Android emulator for PC. It enables emulating applications in APK format (Android application package file), features portrait and landscape mode, makes possible to browsing the Web and provides other specific functions of Android platform [6].
4. Smartphone LG GT540[2] with Android operating system. It has TFT resistive touch screen, which enables usage of an ordinary stylus as the pointing device. Screen size in this smartphone is 3 inches and the resolution is 320x480 pixels. The phone was fixed to the adjustable handle, to ensure proper position and stability.

Fig. 2. YouWave Android emulator and LG GT540 smartphone showing the tested mobile touch screen facebook version

3 Experiments

The aim of our study was to investigate mobile application usability testing using eyetracking. In order to test this method we conducted usability study of social network service facebook.com in its mobile version, using the eyetracking equipment. For this purpose we have prepared several tasks, intended to demonstrate whether the

[2] http://www.lg-optimus.com/

use of this application may evoke problems, and if so what was their cause. The tasks were partly based on the principles of creating usability testing scenarios [5]. Before the experiment, several fictitious user accounts were created and combined in a small network of friends, which enabled better monitoring of user actions in the service. The shortened content of the tasks which were given to users is presented below:

1. Organize your birthday. Create an event scheduled for March 20, at 7 PM in the "Tawerna club". In the description write "no gifts". Invite your two friends to this event.
2. Create fan page site called "Mobile Eyetracking". Change the settings so that it will be only available to people over 18 years of age.
3. You want to call your friend "Peter Eyetracking" but you do not know his phone number. Find it on facebook.
4. Add a comment to your friend's photo and click "like" next to it.
5. Check your recent notifications.

We conducted two series of experiments. One with the real phone placed on the handle, second using YouWave emulator shown on the PC screen. Each of them was carried out with five different users. They all were students of the Wroclaw University of Technology. All of them were quite experienced users of different web applications, but their knowledge and time spent on the facebook service was varied.

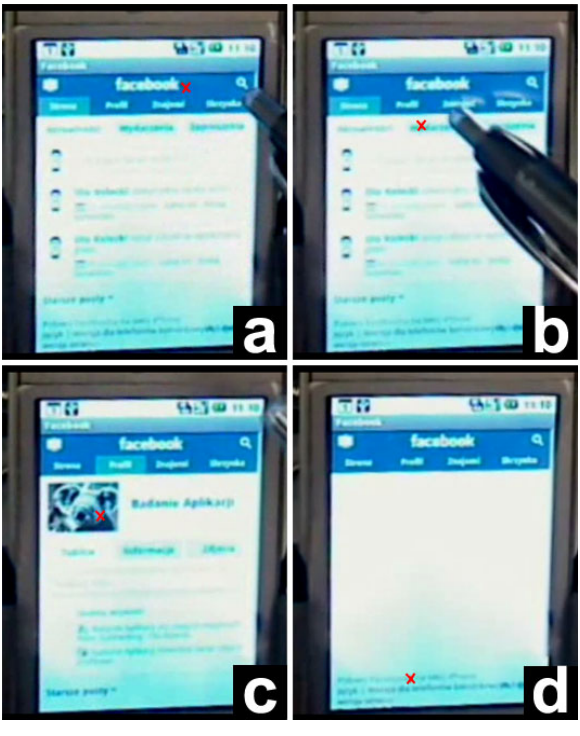


Fig. 3. Sequence of images with a superimposed "X" indicating an actual gaze point of the user: a) notice of the menu element b) pressing the chosen menu element c) brief look on koala bear face d) quick look at the text that shows up

Every participant was wearing the eye tracking HMO headband. Calibration was done for each one of them. The HMO test procedure allows gaze tracking on the PC monitor, so that part progressed as planned. Developing testing environment for eye tracking for mobile devices was one of the challenges which we had to face during researches. We attached a web camera to the headband, so it would record the screen of the smartphone. In the meantime we ran GazeTracker software[3], on the computer screen (over notepad application), calibrated to the size and position of the phone. The eyetracking data (red "X" sign) was than extracted and placed on the images recorded by the webcam (Fig. 3).

After presentation of the test rules each participant was asked to fill out the "before test" questionnaire. Questions concerned identification data, current occupation, education, interests, but also the experience of using a touch phone, mobile applications and knowledge of web page version of facebook.

The test consisted of execution of the five tasks presented above, read by the moderator. During the experiment, time, directories and overall success or failures were recorded. Also notes about specific behavior and mistakes made by users have been taken.

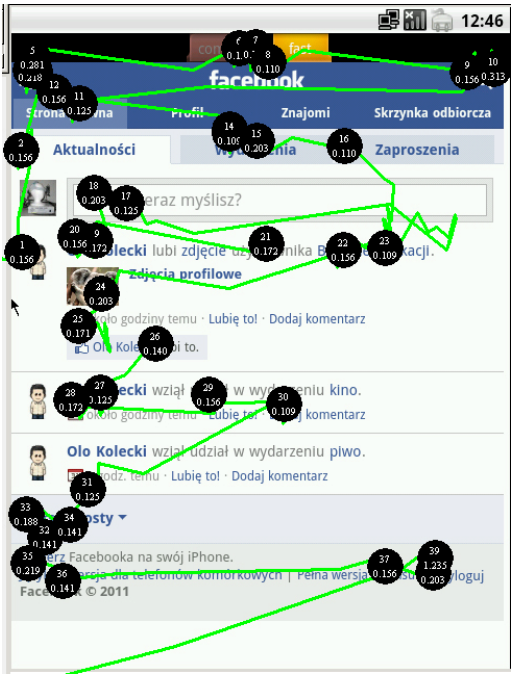


Fig. 4. Example of the sequence of gaze points with fixation times

[3] http://www.asleyetracking.com/Site/Products/SoftwareSolutions

After the experiment users were asked to fill post-test questionnaire concerning their feelings about the test. They also had the possibility to express their thoughts on the tested applications. However we focused our analysis on material obtained from eyetracker software. For emulator test we got all the data provided by GazeTracker software, such as gaze plots (Fig. 4). For smartphone we had video files with imposed gaze points (Fig. 3).

4 Results of the Experiments

The results of both tests are presented in Tables 1 to 4. They present detailed results and comparisons between real phone and emulated facebook application tests.

Table 1. Time results of completed tasks on YouWave Android emulator

	Familiar with Facebook mobile version	Task 1 Time [s]	Task 2 Time [s]	Task 3 Time [s]	Task 4 Time [s]	Task 5 Time [s]
Person 1	Yes	58	85	64	76	11
Person 2	No	124	137	66	145	40
Person 3	No	65	147	19	71	37
Person 4	Yes	60	141	40	50	20
Person 5	Yes	73	47	8	46	5

Table 2. Time results of completed tasks on touch screen smartphone

	Familiar with Facebook mobile version	Task 1 Time [s]	Task 2 Time [s]	Task 3 Time [s]	Task 4 Time [s]	Task 5 Time [s]
Person 6	Yes	35	62	6	28	10
Person 7	Yes	57	68	63	75	35
Person 8	No	94	75	95	80	45
Person 9	No	92	59	20	37	7
Person 10	Yes	41	65	22	46	9

Table 3. Comparison of average times of tasks

	Average time [s]				
	Task 1	Task 2	Task 3	Task 4	Task 5
YouWave Emulator	76	111	55	78	23
Smartphone	64	66	41	53	21

Table 4. Comparison of the test methods for mobile eyetracking testing

	Comparison of the test methods	
	YouWave Emulator	Touch screen smartphone
User comfort	Big screen in front of user, good angular view, very comfortable.	Small screen and interaction using stylus, phone was far from users, so they had small angular view on it, not to comfortable.
Eyetracking data precision	Very precise.	After merging gaze points with recorded images, the data was slightly imprecise.
Calibration	Easy calibration, same as for normal web page usability tests.	Difficult and long calibration, we needed to transform calibration points to the size of the mobile device screen.
Interaction	Users used mouse to interact with the application, so it did not fully simulate the way they would work with the mobile application.	Interaction using stylus, identical to the way most people use smartphones
Other remarks	The keyboard in the emulator was very poor, and users had many problems with typing text using it.	Application was tested in its natural environment.

After the experiments we gathered all the results and observations, and analyzed them. Our tests showed that the tested application has many usability issues and users were responding negatively to it. Most of the problems with using this application could be found using standard user testing without eyetracking, however data obtained from those tests allows us to benefit from all the information that we get from eyetracking testing. We are able to determine where users look while doing specific actions in this application. During the processing of the eyetracking data we came to some interesting results, for example most of the time users were scanning the application with f-shaped pattern [10]. Furthermore generated reports have clearly shown the problems that users had with particular tasks. For example in task nr 5, when users got to the photo, they had troubles with finding the comment box, because it was hidden under the photo (Fig. 5). The times of the task completion has varied between users who were familiar with this application and those who used it for the first time. Experienced users coped with the tasks faster than new ones (Tables 1 and 2). Generally users completed tasks on smartphone faster than on the emulator (Table 3). Such result is probably caused by the method of interaction used in both cases. Performing actions by touching closely aligned buttons on small screen is probably faster than using computer mouse on bigger PC screen.

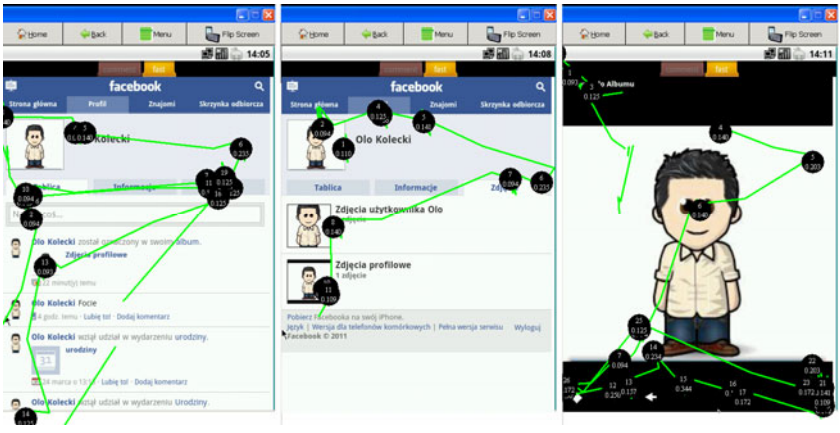


Fig. 5. Task 5 fixation map showing user performing consecutive steps of getting to the profile picture and searching for the comment box

5 Summary

Our experiments have shown that eyetracking is also a very useful method in the mobile devices testing.. In our experiment we used emulator and smartphone application. Both of those methods of interaction had some advantages and disadvantages (Table 4) in the usability testing. Because of that we recommend to perform usability tests using PC emulator and smartphone to get the best results. Smartphones are the target environment for those applications, so we should always use them in the usability tests, however the PC emulator testing together with bigger displays gives better perspective of the application and allows to obtain more visible and clear eyetracking data. Moreover our emulator test was performed on a normal screen with mouse interaction. It would be better if such tests were conducted on a touch screen to make the interaction the same as on mobile device.

We had many troubles with setting the testing environment with standard eyetracker, so to perform fast and effective eyetracking tests on mobile devices it would be recommended to obtain a professional mobile eyetracker with advanced camera and software, that would allow easy and fast calibration and analysis for objects outside of the computer screen. Application developers should also perform usability tests on different smartphones, because they have different parameters such as screen size and buttons.

It is obvious that the mobile application will be more and more popular, so surely mobile applications usability testing will become an important factor, and mostly used web usability methods such as clictracking and eyetracking will be adapted to mobile environments. Moreover we could for example try to analyze the eyetracking data with collaborative intelligence methods, like data mining, to get some interesting dependencies between the users. It could provide us with information about the regions of the application that particular group of users is mostly interested in or create some graphs showing the route for given task for those groups of users. There is still plenty room for research in this area.

Acknowledgements. This work has been partially supported by the Polish Ministry of Science and Higher Education within the European Regional Development Fund, Grant No. POIG.01.03.01-00-008/08.

References

1. Applied Science Group, Eye Tracking System Instructions. ASL Eye-Trac 6000, Pan Tilt/Optics (2006)
2. Applied Science Group, Eye Tracking System Instructions. ASL Eye-Trac 6000 Head Mounted Optics (2006)
3. Chynał, P., Sobecki, J.: Comparison and Analysis of the Eye Pointing Methods and Applications. In: Pan, J.-S., Chen, S.-M., Nguyen, N.T. (eds.) ICCCI 2010. LNCS, vol. 6421, pp. 30–38. Springer, Heidelberg (2010)
4. Duchowski, A.T.: Eye tracking methodology: Theory and practice, pp. 205–300. Springer-Verlag Ltd., London (2003)
5. Go, K.: What Properties Make Scenarios Useful in Design for Usability? In: Kurosu, M. (ed.) HCD 2009. LNCS, vol. 5619, pp. 193–201. Springer, Heidelberg (2009)
6. Information about YouWave Android project, http://www.youwave.com/
7. Information about growth of smartphone market, http://www.comscore.com/Press_Events/Press_Releases/2010/9/European_Smartphone_Market_Grows_41_Percent_in_Past_Year
8. International Standard ISO 9241-11. Ergonomic requirements for office work with visual display terminals (VDTs) – Part 11: Guidance on Usability. ISO (1997), http://www.cs.tufts.edu/~jacob/papers/barfield.pdf (March 13, 2011)
9. Mohamed, A.O., Perreira Da Silva, M., Courbolay, V.: A history of eye gaze tracking (2007), http://hal.archivesouvertes.fr/docs/00/21/59/67/PDF/Rapport_interne_1.pdf (March 10, 2011)
10. Nielsen, J.: Jakob Nielsen's Alertbox, F-Shaped Pattern For Reading Web Content (April 17, 2006), http://www.useit.com/alertbox/reading_pattern.html (March 12, 2011)
11. Nielsen, J.: Jakob Nielsen's Alertbox, Mobile Usability (July 20, 2009), http://www.useit.com/alertbox/mobile-usability.html (March 10, 2011)
12. Wesson, J.L., Singh, A., van Tonder, B.: Can Adaptive Interfaces Improve the Usability of Mobile Applications? In: Forbrig, P., Paternó, F., Mark Pejtersen, A. (eds.) HCIS 2010. IFIP AICT, vol. 332, pp. 187–198. Springer, Heidelberg (2010)

An Approach for Improving Thai Text Entry on Touch Screen Mobile Devices Based on Bivariate Normal Distribution and Probabilistic Language Model

Thitiya Phanchaipetch and Cholwich Nattee

School of Information, Computer, and Communication Technology,
Sirindhorn International Institute of Technology, Thammasat University
thitiya.phanchaipetch@student.siit.tu.ac.th, cholwich@siit.tu.ac.th

Abstract. This paper presents an approach to improve the correctness for Thai text inputting via virtual keyboard on touch screen mobile phones. The proposed approach is to generate candidate character based on statistical model from bivariate analysis of pre-collected coordinate data and apply the character trigram model to each candidate character sequence. From user's touch positions, a set of candidate characters with high position-based probability is generated. Then, the character trigram model is applied to each generated candidate characters sequence. For each character sequence, a probability is computed from the weighted combination of position-based and character trigram models. In the end, the character sequence with the highest probability is selected to be the most appropriate sequence. Experiments were conducted to compare the typing accuracy between an ordinary Thai virtual keyboard and our proposed algorithm using the same Thai keyboard layout. Results demonstrate that the proposed algorithm provides the improvement in the text entry accuracy in both character levels and word levels.

Keywords: virtual keyboard, thai text entry, touch screen mobile phone, touch screen mobile device, trigram model, thai keyboard, bivariate normal distribution.

1 Introduction

The majority of the worlds digital experiences now happen through mobile devices especially on touch screen mobile phones. One trend in consumer electronics is touch-sensitive screen that are controlled by human finger motions rather than a cursor or pointing device. Smartphone sales to end users were up 72.1% from 2009 and accounted for 19% of total mobile communications device sales in 2010. [1]

Virtual keyboard is a software component that allows a user to enter characters and can usually be operated with touch screen device where data entry is required a physical and keyboard is not available. Focusing on Thai language,

J.-S. Pan, S.-M. Chen, N.T. Nguyen (Eds.): ACIIDS 2012, Part III, LNAI 7198, pp. 187–196, 2012.

there are more alphabets in Thai than in English. The small area of touch screen mobile device causes the problem that each key on keyboard layout is located very close to each other and the size of each button becomes small. Even though every standard Thai keyboard is splitted into 2 sub-layouts which are shift sub-layout and non-shift sub-layout to cover all Thai consonants and vowels, it is still hard to accurately touch on the intended key button. Most human fingers are always larger than key buttons. Pressing on a small button with a large finger tends to generate error, especially when virtual keyboard has no tactile feedback. This problem is widely known and it decreases efficiency of Thai text entry on touch screen mobile phone.

This paper focuses on improving Thai key prediction for each press to reduce the typing error. Based on the assumption that each user has his/her own characteristic of typing. He/She will touch on particular position for the same key button. Each time, user's touch point is located closer to the set of points they have typed. Individual characteristic for each user can be formed. Bivariate Normal Distribution is assumed to be a statistical model for each particular key. From this statistical data, probability that user wants to type which key button given by the currently user touch point can be found. The right character should be the key which has the highest probability. We obtain the set of candidate keys which have high probability among all characters in one sub-layout. Another assumption in this paper is that most of words that user types are common words widely used in various articles. After obtaining the set of candidate keys, we concatenate all candidate keys to form candidate character sequence in every possible way. So, we grade the candidate sequence with probabilistic language model trained from a Thai corpus in order to find out the best candidate sequence with highest probability and display to user.

This paper consists of five parts. Background of this paper and other related works is explained in the next section. Bivariate Analysis-based Candidate Generation, Probabilistic Language Model for Assigning to candidate and Candidate Ranking are described in Section 3. Section 4 describes experiment setting and results and the last section is conclusion of the paper and our future direction.

2 Background and Related Works

There are many ways to improve performance of text entry on touch screen devices. Significantly changed in keyboard layout is one way to improve typing performance. Dvorak layout is well known as a QWERTY alternative, patented in 1936 by Dvorak and Dealey[7] The advantage of this layout is it uses less finger motion, increases typing rate and reduces errors compared to the standard QWERTY keyboard. MacKanzie, Zhang & Soukoreff (1999)[3] introduced a new virtual keyboard where the letters were laid in alphabetic ordering in two columns. The problem of new layout is it quite hard and takes time to make user get familiar with. This paper focuses on improving text entry base on standard keyboard layout, Thai Kedmanee which conforms to the Thai keyboard layout on Microsoft Windows to make user get more familiar with.

Research on improving input methods on standard keyboard layout has grown over recent year due to the limitation size of screen and lack of tactile input feedback. Potipiti et al.[5] introduced the approach by applied trigram model and error correction rules for intelligent Thai key prediction and English-Thai language identification. First technique is language identification which performs language switching automatically between Thai and English without pressing language-switch button. The second technique is Thai key prediction which applied character trigram probabilistic model and compare probability between Thai character sequence with shift key and without shift key. Trigram is well-known probabilistic language model that applied in this method. It proved that trigram is work for predicting the next character in such a text sequence. So, we try to apply this technique to our proposed method.

MacKanzie and Zhang[4] introduced Eye typing using word and letter predication in English language. User types by gazing at on-screen keyboard using eye tracker and software. Fixation algorithm is applied to determine which button on the keyboard receives eye-over highlighting. They considered these keys as candidate key. Word prediction was introduced by language model and a list of candidate word is produced as entry proceeds. The user selects the desired word. Even though these techniques can reduce the number of keystrokes per character of text entered and may increase text entry speed, it was sensitive to the correctness of the first several letters in a word and works only in letter prediction mode.

Janpinijrut et al.[2] introduced an approach for improving Thai text entry on standard Thai Keyboard based on distance and statistical language model. Touch area is defined by the area nearby the touch point. The characters which locate within touch area are considered as candidate characters. Each candidate characters are weighted based on their distances from the touch points. After that, the character trigram model is applied to select the combination of characters with the highest probability and suggest to the user.

After the experiments from this technique, we found that most of individual user generate the same typing error many times. This technique shows improvement in term of characters but it is not fit with variety of user typing characteristics because each touch point predicts character based on distance, not related to statistical analysis. The method that introduced in this paper is the alternative way to improve Thai text entry by consider individual typing characteristic and create adaptive user keyboard.

3 The Proposed Approach

The proposed approach consists of three main steps as shown in Figure 1, i.e., candidate generation based on bivariate analysis, probabilistic language model for assigning to candidate and candidate ranking.

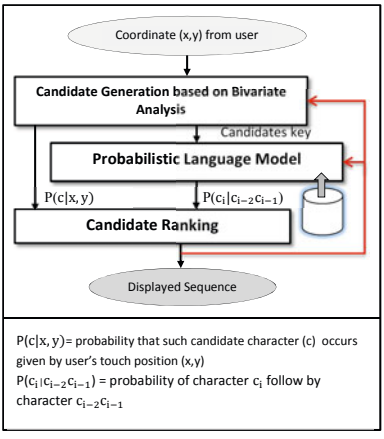


Fig. 1. Overall of proposed algorithm

3.1 Candidate Generation Based on Bivariate Analysis

From our assumption, suitable touch area should be the area that most people touch on. We assume each key has its own probability distribution as Bivariate Normal Distribution. In this case, we interested in the joint occurrence and distribution of values of the independent and dependent variable together. Bivariate Normal Distribution is joint distribution of two variables which are x and y. These values are considered to be the coordinates for a point geometry and can be obtained when user touches on screen. So, we find out which area those users mostly touch on each character. Coordinate, x and y are collected by shuffling all of Thai characters that exist on standard Thai keyboard. It composed of 43 characters per sub-layout. There are two sub-layouts which are shift and non-shift and each are mixed with alphabets, consonants and special characters. Both sub-layouts have the same layout. Therefore, it has 43 unique keys. 5 users are asked to type all 43 distinct characters 3 times in free style. After that, we obtained 15 coordinates of each character position. These coordinates are used through all step of proposed algorithm.

Figure 2 shows standard keyboard layout with distribution of 43 distinct character positions.

The formula (1) is known as Bivariate Normal Probability Density Function. Suppose we have two random variables, coordinate x and y. This formula can be integrated to obtain the probability that random coordinate x and y take a value in a given pre-collected coordinate data set for each character. Bivariate Normal Density Function is used to find the point of Normal Distribution curve:

Fig. 2. Distribution of touch point of each character position for non-shift sub-layout

$$P_{pos}(c_i|p_i) = \frac{1}{2\pi\sqrt{\sigma_x\sigma_y(1-\rho_{xy}^2)}} \times exp\left\{-\frac{1}{2(1-\rho_{xy}^2)}\left[\left(\frac{p_{ix}-\mu_x}{\sqrt{\sigma_x}}\right)^2 + \left(\frac{p_{iy}-\mu_y}{\sqrt{\sigma_y}}\right)^2 - 2\rho_{xy}\left(\frac{p_{ix}-\mu_x}{\sqrt{\sigma_x}}\right)\left(\frac{p_{iy}-\mu_y}{\sqrt{\sigma_y}}\right)\right]\right\} \tag{1}$$

$P_{pos}(c_i|p_i)$ is probability that
candidate character c_i occurs given by user's touch position p_{ix} and p_{iy}.
It is involving with the individual parameter μ_x, μ_y, σ_x, σ_y and ρ_{xy} :
σ_x is covariance of a set of coordinate x belongs to that character,
σ_y is covariance of a set of coordinate x belongs to that character,
μ_x is mean of a set of coordinate x belongs to that character,
μ_y is mean of a set of coordinate y belongs to that character,
ρ_{xy} is ccorrelation coefficient between 2 values, coordinate x and y that belongs to that character.

Mean value can be calculated from below formula:

$$E(x) = \frac{1}{N}\sum_{i=1}^{N} x_i \tag{2}$$

Covariance can be calculated from below formula:

$$Var(x) = \sum_{i=1}^{N} \frac{(x_i-\mu)^2}{N} \tag{3}$$

The correlation coefficient is computed as:

$$Corr(x,y) = \frac{\sum_{i=1}^{N}(x_i-\mu_x)(y_i-\mu_y)}{\sqrt{\sum_{i=1}^{N}(x_i-\mu_x)^2\sum_{i=1}^{N}(y_i-\mu_y)^2}} \tag{4}$$

where

N is a number of data in one data set,
μ is mean of the data set.

When user touches on the screen to type a character, coordinate input x and y from user are calculated using the formula shown in Equation (1) with all 43 coordinates data set to find the probability of each character given by user coordinate input. After that, we can find a number of characters which have highest probability among 43 characters. In this paper, we conduct an experiment by select only top 3 characters to be candidate keys and process in the next step.

3.2 Probabilistic Language Model for Assigning to Candidate

An n-gram model is a type of probabilistic model for predicting the next item in such a sequence. This technique is applied to propose method to predict the next character given the previous character that user has typed before. In this paper, character trigram model is used in order to rank candidate key because mobile device has limited resource.

After generating candidate keys from previous step, each candidate concatenates all possible candidates to create all possible character sequences called candidate sequence. Then, we calculate probability for each candidate sequence by using character trigram model.

$$P_{lang}(c_1 c_2 \dots c_n) = \prod_{i=1}^{N} P(c_i | c_{i-2} c_{i-1}) \tag{5}$$

where $P_{lang}(c_1|c_2 ... c_n)$ is probability of each candidate sequence, $P(c_i|c_{i-2}c_{i-1})$ is probability of character c_i follow by character $c_{i-2}c_{i-1}$

From this point, we obtain a set of candidate sequences. Each is assigned by the probability from language model.

3.3 Candidate Ranking

The last step of proposed method is Candidate Ranking. From this point, each candidate sequence has its own probability given by position from 3.1 and given by language model from 3.2.

Candidate sequences are ranked by following scheme:

$$P(c_1 c_2 \dots c_n | p_1 p_2 \dots p_n) = \arg\max_{c_1 c_2 \dots c_n} \big(\alpha \cdot \sum_{i=1}^{n} P_{pos}(c_i|p_i) + \beta \cdot P_{lang}(c_1 c_2 \dots c_n) \big) \tag{6}$$

where

$P_{pos}(c_i|p_i)$ is probability that such candidate character c_i occurs given by user's touch position p_i which is calculated from section 3.1,

$P_{lang}(c_1c_2 \ldots c_n)$ is probability that such candidate sequence $c_1c_2 \ldots c_n$ occurs given by language model which is calculated from section 3.2 α and β are weight to counterbalance between character prediction by position and probabilistic language model. These two values should greater than 0 and sum of these two values is required to be exactly 1.0.

In section 5, we conduct experiment to find the most suitable value of α and β and the result shows that the best appropriate value of α is 0.6 and 0.4 for β.

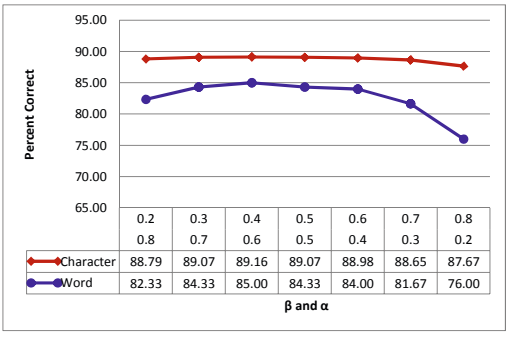
Each candidate sequence is processed by (6) scheme and rank. The candidate sequence which is in the first rank is suggested and displayed to user. This gives a result that every time user types a character, there is a chance that displayed sequence has changed to right sequence, even though they type wrong character. This is because ranking is processed every time user types a character. From assumption, the most of words users type are common words and widely used in various document. Even if users type position is not exactly on intended key button, the right character sequence should have highest probability in term of language consideration. Sometimes the right sequence is ranked to the top after processed and changes from wrong sequence to the right one.

4 Experiments and Results

In order to evaluate the performance of the proposed method, we conduct experiments based on collected coordinates set from user and compare the performance between proposed method and ordinary method.

There are many touch screen phones available in the market. Android is quickly becoming one of the most popular tools for mobile application development. The most prominent is Android's open-source nature. HTC Desire HD is capacitive touch screen Android phone that we use in our work. Its resolution is 480×800 pixels. Android 2.2.1 operating system is installed. CN Thai keyboard layout is chosen to be experiment keyboard. Experiment is fixing to portrait orientation because it has smaller keyboard area than landscape orientation. All experiments are implemented and tested on this phone.

For the experiment, we first find the most suitable value of α and β before we can evaluate the actual efficiency of proposed method. 5 million words from BEST 2010[6] Thai corpus created by NECTEC is used to train and create a character level 3-gram model. We randomly choose 30 words from the corpus to be test set. Each set consists of 10 words that occur with high, medium and low frequency in corpus as well as long, medium and short word length in the same average number. We let 10 users type all words in 3 test sets word by word in their style on mobile phone. Then we obtain 30 coordinates per user. Every coordinate is processed by the proposed algorithm on the mobile phone. We match these 30 output characters sequences against correct words to find the accuracy in character and word level. In character level, we match and count the number of correct characters in words. In word level, we match word by word.



β and α	0.2 / 0.8	0.3 / 0.7	0.4 / 0.6	0.5 / 0.5	0.6 / 0.4	0.7 / 0.3	0.8 / 0.2
Character	88.79	89.07	89.16	89.07	88.98	88.65	87.67
Word	82.33	84.33	85.00	84.33	84.00	81.67	76.00

Fig. 3. Percent correctness in character and world level of various value of α and β

Figure 3 shows the accuracies at the different value of α and β in character and word level.

The graph shows that the correctness is increasing if α is decreased and β is increased. It obviously shows that probabilistic language model in section 3.2 takes a role in improvement. But when β is increased to 0.5, the correctness is continuing take a fall and dramatically dropped when β is 0.8. This is because of the lack of known word list. So, the unknown word can be the output if we set β to high value. We plan to apply dictionary to improve correctness in the future work. From this point, the most suitable values for and are 0.6 and 0.4 respectively which yield around 89.16% in character level and 85% in word level.

From the experimental results, the performance in word level is slightly dropped when we compare to character level because the method did not apply any statistics or information related to words. Character trigram model is used only in the proposed method.

After we obtained the most suitable values of α and β, 10 users are asked to type all 30 words 3 times. Five of them are the same users as we kept pre-collected coordinate data from them. The details had described in section 3.1.

Finally, we process coordinates from all users by using our proposed method and ordinary method to find the actual performance of proposed method against the ordinary typing method. The correctness of the ordinary method in character level is 96.65% and 80.67% in world level. While the correctness of our proposed method in character level is 98.91% and 93.78% in word level. This result shows that our method provides the improvement in the text entry accuracy on touch screen mobile phone for both character and word level. To ensure that the result of proposed algorithm is not better than ordinary algorithm *by chance*, we also conduct the paired t-test which assesses whether the means of two groups are statistically different from each other. The score associated with t-test is 1.69152×10^{-5} in word level and 2.69239×10^{-5} in character level. These values are less than 0.05. Therefore, there is a significant difference between of two groups.

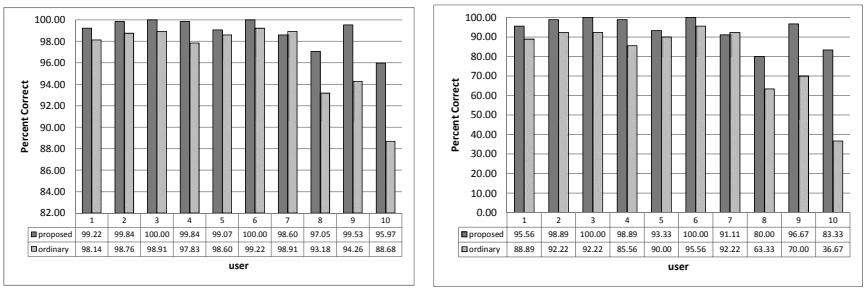


Fig. 4. Percent correctness in character and word level of 10 users by using two algorithms respectively

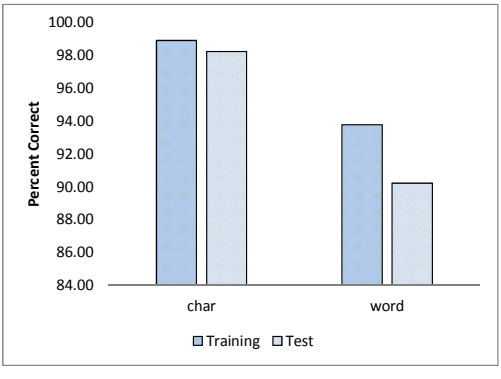


Fig. 5. Percent correctness in character and world level of various value of α and β

Moreover, based on the assumption that each user has their own characteristic of typing and can be formed to create distribution. If we focus on our algorithm by separate user into two groups, first group is five users that we kept pre-collected coordinate data from them (Training group) and second group; another five users (Test group), the performance for first group is better than second group in both character and word level. It ensures this assumption is correct because pre-collected coordinate that we obtained from first group is used to create bivariate normal distribution for each key. So, when user in training group types characters, the probability that such character occurs given by users touch position is probably higher than test group because the probability is based on distribution which generated from training group. Figure 4. shows percent of correctness of training group and test group in character level and world level.

4.1 Conclusion and Future Works

This paper has proposed a new approach to improve Thai text entry on touch screen mobile devices. Bivariate normal density on touch positions is used to

generate candidate keys and then processes by using character trigram model. After that, all candidate sequences are ranked based on probability given by users coordinate and probability given by language model and suggested to user. We conduct experiment to find the best value of and to weight to counterbalance between character prediction by position and probabilistic language model. The most suitable value of α is 0.6 and β is 0.4. Our algorithm performs better than the ordinary method in both character level and word level. The correctness of the ordinary method in character level is 96.65% and 80.67% in world level. While the correctness of our proposed method in character level is 98.91% and 93.78% in word level.

For the future works, we plan to use list of frequently used words in order to improve performance of Probabilistic Language Model for Assigning to candidate sequence. Furthermore, we aim to do word completion and create user adaptive keyboard.

Acknowledgments. This research is financially supported by Thailand Advanced Institute of Science and Technology-Tokyo Institute of Technology (TAIST & Tokyo Tech), National Science and Technology Development Agency (NSTDA) Tokyo Institute of Technology (Tokyo Tech), Sirindhorn International Institute of Technology (SIIT), Thammasat University (TU) and Telecommunications Research and Industrial Development Institute (TRIDI).

References

1. Gartner, Inc.: Gartner Says Worldwide Mobile Device Sales to End Users Reached 1.6 Billion Units in 2010; Smartphone Sales Grew 72 Percent in 2010 (2010), http://www.gartner.com/it/page.jsp?id=1543014
2. Janpinijrut, S., Nattee, C., Kayasith, P.: An Approach for Improving Thai Text Entry on Touch Screen Mobile Phones Based on Distance and Statistical Language Model. In: Theeramunkong, T., Kunifuji, S., Sornlertlamvanich, V., Nattee, C. (eds.) KICSS 2010. LNCS, vol. 6746, pp. 44–55. Springer, Heidelberg (2011)
3. Mackenzie, I.S., Zhang, S.X., Soukoreff, R.W.: Text entry using soft keyboards. Behaviour and Information Technology 18, 235–244 (1999)
4. Mackenzie, I.S., Zhang, X.: Eye typing using word and letter prediction and a fixation algorithm
5. Potipiti, T., Sornlertlamvanich, V., Thanadkran, K.: Towards an intelligent multilingual keyboard system
6. Thailand National Electronics and Computer Technology Center (NECTEC): Benchmark for Enhancing the Standard of Thai language processing 2010, BEST 2010 (2010), http://www.hlt.nectec.or.th/best/?q=node/10
7. West, L.J.: The standard and dvorak keyboards revisited: Direct measures of speed. Tech. rep., Measures of Speed, Santa Fe Institute Working Papers Paper (1998)

An Iterative Stemmer for Tamil Language

Vivek Anandan Ramachandran[1] and Ilango Krishnamurthi[2]

Department of Computer Science and Engineering, Sri Krishna College of Engineering and Technology, Coimbatore – 641 008, Tamilnadu, India
{rvivekanandan,ilango.krishnamurthi}@gmail.com

Abstract. Stemming algorithm is a procedure that attempts to map all the derived forms of a word to a single root, the stem. It is widely used in various applications with the main motive of enhancing the recall factor. Apart from English, researches on developing stemmers for both the native and the regional languages are also being carried out. In this paper, we present a stemmer for Tamil, a Dravidian language. Our stemmer effectiveness is 84.32%.

Keywords: Information Retrieval, Stemming, Tamil.

1 Introduction

A program that performs stemming is referred as a stemmer [1]. Stemmer attempts to map a derived form of a word to its root. For example, stemmer maps the word *creation* to the term *cre*. Resultant of a stemmer need not be a proper meaningful word. This could be understood from the example as *cre* is not a meaningful word.

Stemmers are widely used in the query based systems such as web search engine, question answering etc as a factor to retrieve more number of documents relevant to the input . This is mainly due to the reason that such systems treat the word and its derived forms as one and the same. For example, when a user wants to search the documents with the term *creating* they might also need the documents with the term *creates* or *created*.

In recent years web documents and other information are published in languages other than English too. These published information made researchers to focus on the need for developing computational supportive tools such as stemmer, lemmatizer, parts-of-speech tagger etc for languages other than English. In this paper, we discuss our experience in developing stemmer for Tamil[1][2]. In this paper, we propose a suffix stripping stemmer for Tamil.

This paper is organized as follows: In Section 2 we discuss the researches related to our stemmer. In Section 3 we brief about the Tamil suffixes. The difficulties in developing stemmer for Tamil language are highlighted in Section 4. In Section 5 we describe about our algorithm. In Sections 6 and 7, we present the evaluation analysis and the concluding remarks respectively.

[1] Tamil is a Dravidian language. It is spoken majorly by the Tamil people of the Southern India and Substantial minorities in Malaysia, Mauritius and Vietnam. Throughout the paper transliterated form of Tamil as quoted in the Appendix Section is used.

J.-S. Pan, S.-M. Chen, N.T. Nguyen (Eds.): ACIIDS 2012, Part III, LNAI 7198, pp. 197–205, 2012.

2 Related Work

Earlier, stemmers were primarily developed for English language[3][4]. But later due to the corpus growth of languages other than English, there was an increased demand from the research community to develop stemmers for other languages too. In the case of Indian languages, stemming was first reported for Hindi in 2003[5]. Slowly investigations for other languages such as Bengali[6], Urdu[7], Malayalam[8] and Punjabi [9]were also carried out. However, there is no readily available stemmer for Tamil. In this paper, we present our research experiences in developing a Tamil Stemmer.

To develop a stemmer for Tamil or any other languages, a basic approach to carry out the process is required. The most common approaches used for developing a stemmer are Brute force, Affix Stripping, N-Gram, Hidden Markov Model (HMM), Corpus based technique, Clustering method, Finite-State-Automata method, Morphological process, String distance measure, Hybrid approaches. Among all the existing approaches, we make use of affix stripping because of its inherent support to develop a stemming algorithm in an easier and faster way.

Most of the existing stemmers remove the suffixes based on the longest matching word. For example among the matching suffixes *ates*, *tes*, *es* and *s* existing in the word creates, the suffix *ates* will be removed by a stemmer. Similar to most of the existing stemmers, we remove the suffixes based on the longest matching word.

To develop a stemmer for a language, a preliminary study on the possible suffixes of a word in the corresponding language need to be taken. In the next Section, we explain about the possible suffixes for Tamil.

3 Tamil Suffixes

In Tamil, a word usually contains a root to which one or more affixes can be attached. The affixes can be either a prefix or a suffix. We have designed our algorithm to handle only Tamil suffixes, so we discuss only about it. Tamil word can have multiple suffixes there is no exact limit for attaching the number of suffixes to a Tamil word.

To know about the possible suffixes that a word in Tamil language one can refer to the flow charts [10] and [11]. Besides considering the possible suffixes forms of a word, a stemming algorithm has to consider certain standard computing issues. In the next Section, we explain the computing issues considered for developing our stemmer.

4 Computing Issues

Developing a Tamil stemmer is not a straightforward task. In this Section, we explain the major difficulties faced by us while designing the algorithm. They are briefed as follows:

4.1 Homographs

Homographs are the words that have identical pronunciations but different meanings. In Tamil, there are numerous instances of homographs. For example, the word *aaNTavan* denotes either the noun *God* or the verb *ruled by a male* formed from the root *aaL*. It is difficult for a rule-based stemmer to map such terms to their root. Hence, we decided to frame our algorithm by do not considering the homograph issues.

4.2 Irregular Verbs

Irregular verbs do not follow standard patterns in their tense form. For example, the past tense of the verb *say* is *said* and not *sayed*. Mapping such forms of word to a single root is a difficult task. Such cases exist in Tamil also. For example, the past, present and future tense of the verb *sol* (*say*) are *sonneen* (*I said*), *solkiReen* (*I am saying*) and *solveen* (*I will say*) respectively. Following the standard patterns the past tense for *sol* should be *solneen*. However, this is not correct. Devising rules to handle such case is arduous as it needs a deep look up dictionary. Hence, we decided to consider this issue in the future version.

4.3 Proper Noun Derivations

A proper noun usually indicates a particular thing. In certain cases, proper noun end letters match with the normal suffixes. Assuming those patterns to be suffixes, most of the stemmers remove them from the proper noun. For example, Porter stemmer maps the proper noun *creator* to *cre* due to the assumption that *ator* is a suffix. To overcome such cases, it is very difficult to realize a proper noun by framing hand-crafted rules. Therefore, similar to most of the algorithms, if the common suffix patterns exist in a proper noun we decide to stem it. For example, our approach maps *iyakkiyavan* (*A person who operated*) to *iyakki* (*operated*).

4.4 Handling Non Derived Words Ending with Usual Suffix Pattern

In some cases, word ends with few patterns that match with a suffix but that pattern does not denote a suffix. For example:

- In English, the word *ring* contains *ing* which usually denotes a suffix but not in this case.
- In Tamil, the word *kathai* (*story*) contains *ai* which usually denotes a suffix.

It could be inferred that to handle this case a stemmer needs a heavy look-up table and this table cannot be constructed easily. So we decided to handle this case in the future versions.

4.5 Study on the Number of Iterations Needed to Remove Suffixes

We have already mentioned that Tamil words can have multiple suffixes. It is discussed in the previous Sections that the best way for removing multiple suffixes is iterating the suffixes in the descending order of their length and removing the suffixes in the input. But an interrogation arises on the following issue:

- For a stemmer to remove n suffixes in an input whether n iterations are needed or it could be done with less than n iterations?

For addressing this issue, we discuss three cases:

4.5.1 For Removing n Suffixes Less than n-1 Iterations Are Needed

Consider the word *nuulkaLiliruntu* (*from the book*). It has three suffixes to be removed viz. *kaL*, *il* and *iruntu*. In our rule-base three suffixes will be in the order *iruntu*, *kaL* and *il*. However, in the example they are in the removal order *iruntu*, *il* and *kaL*. So if our approach follows linear methodology the suffix *iruntu* and *il* matches with the example. Therefore, they are stripped from *nuulkaLiliruntu*. So, the resultant output is *nuulkaL*. However, the suffix *kaL* is not removed from the input. During the next iteration the suffix *kaL* will be removed. So for removing 3 suffixes 2 iterations are needed.

4.5.2 For Removing n Suffixes n Iterations Are Needed

Consider the word *nuulkaLil* (*in the book*). It has two suffixes to be removed *kaL* and. In our rule-base two suffixes will be in the order *kaL* and *il*. However, in the example they are in the removal order *il* and *kaL*. The first iteration removes the suffix *il* and the second iteration removes the suffix *kaL*. So for removing 2 suffixes 2 iterations are needed.

4.5.3 Ascending and Descending Order of Removal of Suffix Affecting the Number of Iterations

Consider the word *nuulkaLukkupatil* (*from the book*). It has three suffixes to be removed viz. *kaL*, *ukku* and *patil*. Our rule-base apart from the three suffixes *kaL*, *ukku* and *patil* will also contain the suffixes *il* and *kku* which matches with the input. They will be in the order *patil*, *ukku*, *kku*, *kaL* and *il*. Let us see what happens when suffixes are iterated in ascending and descending order of the length.

- If suffixes are iterated in descending order of the length the suffix *patil*, *ukku* and *kaL* will be removed in subsequent iterations and the final output will be *nuul* which is the expected one.
- If suffixes are iterated in ascending order of the length only the suffix *il* will be removed and the final output will be *nuulkaLukkupat.* But the desired output is *nuul.*

After analyzing the above three cases for an effective stemmer output we decided to perform the following functionalities in our stemmer:

- Remove *n* suffixes in *n* iterations.
- Iterate the suffixes in the descending order of their length.

We also propose a novel single iteration approach for removing n suffixes in a single iteration.

4.6 Handling Agglutinative Case

Tamil is an agglutinative language; a compound word can be formed from two or more simple words without changing the meaning of the simple words. For example consider the word *maJainiir* (Rain Water) formed from two simple words *maJai* (Rain) and *niir* (Water). Consider the word *maJainiiri_n* (of the rain water). The word is a derived form of *maJai*. Mapping the word *maJainiiri_n* to the simple word *maJai* is a laborious task. So we decided to neglect mapping compound word to simple word.

Apart from the above discussed computational issues, proper computational steps should also be designed to develop a good stemmer. In the next Section, we explain the design portion of our stemmers.

Table 1. Stemming Algorithm

Input	Tamil String (*Input*), Suffix List (*SL*)
Output	Root of the *Input* (*Output*)
Prerequisite	The *SL* should be stored in descending order of suffixes length

```
`Function String stem (Input)
Begin
          1. String Output,
          2. String Temp-Output= ruleBase(Input)
          3. While Input != Temp-Output
                a. Temp-Output = ruleBase(Input)
                b. Input = Temp-Output
          4. return Output
End
Function String ruleBase (temp)
Begin
        1.    Flag = true
        2.    While (Flag)
                a.    Iterate all the suffix one by one
                      i.     If the temp ends with any suffix (say Sf )
                                A.   temp = temp - Sf
                                B.   break
                b.   return temp
End
```

5 Design

Generally, for removing multiple suffixes existing in a word of any language iterative stemmer is used. An iterative stemmer starting from the end of the inflected input word will remove a longest matching suffix at a time and progress towards the root. As discussed in our design section we have designed our stemmer to remove *n* suffixes in *n* iterations. The algorithm pseudo-code is presented in Table 1. The list of suffixes that our stemmer can handle is listed in Table 2.

Consider the input derived word *choRkaLi_nil* (*in the words*) derived from the word *chol*(*word*). During the first , second and third iterations suffixes *il* , *i_n* and *kaL* will be removed respectively. The output will be *choR*. Although the algorithm for stemming Tamil words is designed successfully, it has to be evaluated. In the following Section, we present the analysis carried out by us to study the algorithm's effectiveness.

Table 2. Tamil Suffixes

etirttaaRpool	*appuRam*	*tavira*	*teRku*	*uTa_n*	*oTTi*	*kiR*
aTuttaaRpool	*koNTiru*	*muulam*	*aa_na*	*iyal*	*iTam*	*een*
veeNTiyiru	*veeNTum*	*piRaku*	*tolai*	*avaL*	*kiiJ*	*aaL*
uNkaLee_n	*etirkku*	*pi_npu*	*ava_n*	*mu_n*	*ttal*	*uL*
vaJiyaaka	*aayiRRu*	*pakkam*	*illai*	*avai*	*tiir*	*um*
varaikkum	*veLiyil*	*umee_n*	*poola*	*avar*	*paar*	*tt*
veeNTivaa	*kuRittu*	*meelee*	*aakum*	*a_na*	*atu*	*il*
mu_n_naal	*maatiri*	*kki_nR*	*kiTTa*	*paTi*	*aam*	*pp*
patilaaka	*paarttu*	*taaNTi*	*ki_nR*	*viTa*	*aar*	*ai*
tavirttu	*naTuvil*	*uTaiya*	*pooTu*	*meel*	*aay*	*al*
illaamal	*vaTakku*	*etiree*	*oJiya*	*kiJi*	*kka*	*ya*
veeNTaam	*meeRku*	*uNkaL*	*paNNu*	*ukku*	*iir*	*nt*
kuRukkee	*kuuTum*	*aarkaL*	*koNTu*	*viTu*	*kaL*	*a*
allaamal	*aTiyil*	*vaittu*	*aNTai*	*chey*	*oom*	*p*
varaiyil	*etiril*	*kiiJee*	*uLLee*	*kiTa*	*poo*	*t*
kuuTaatu	*aaTTam*	*arukee*	*taLLu*	*muTi*	*vaa*	*u*
veLiyee	*appaal*	*iirkaL*	*paRRi*	*ooTu*	*i_n*	*v*
iTaiyil	*iIruntu*	*kaaTTu*	*pinti*	*koTu*	*tal*	
aakaatu	*chuRRi*	*nookki*	*patil*	*kkiR*	*vai*	
kiJakku	*arukil*	*viTTu*	*mutal*	*aa_n*	*iru*	

6 Evaluation

In general, evaluation signifies the act of assessing something. To evaluate our stemmer we implemented our algorithm in Java. A sample screen shot of our stemmer is shown in Figure 1.

After developing any stemmer it is important to analyze its capability, i.e., to assess the range of the words that the system is able to stem properly. We requested two students, none of whom are directly or indirectly involved with our project to generate the corpus. They developed a Tamil corpus containing 36720 words derived from 765 roots. The corpus is framed from different portions of Tamil newspapers confining to various domains such as Business, Classifieds, Entertainment, Politics and Sport.

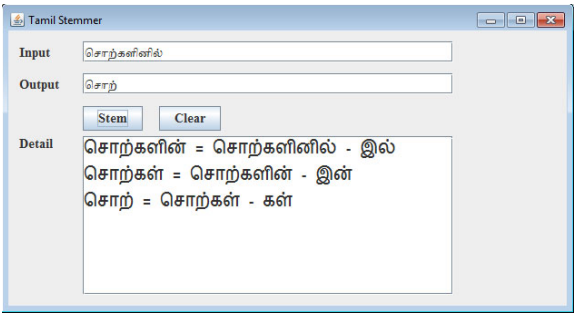


Fig. 1. Sample screen shot of the Tamil Stemmer

It is important to remember as stated in the Introduction that the output of a stemmer need not be a proper linguistic word. Therefore, correctness of a stemmer does not denote the linguistic correctness. A stemmer is said to be accurate if it conforms to the following conditions:

- If it maps all the derived forms of a word to a single root.
- The words mapped by it to a single stem are genuine linguistic variants.
- If it does not stem a non-suffix from a word.

If a stemmer does not map all the considered derived forms of word to a single stem then the phenomenon is called Understemming. An instance of Understemming is a stemmer conflating *tried* to *tri* and *try* to *try* instead of mapping both to *try*. Our stemmers map *cholvava_n* (*A man who is saying*) to *chol* (*say*) and *cho_n_nava_n* (*A man who said*) to *cho_n_n* instead of mapping both to *chol*.

If a stemmer maps the words to a single stem that are genuinely linguistic invariants then the phenomenon is called Overstemming. An instance of Overstemming is a stemmer conflating both the words *cares* and *cars* to *car*, instead of mapping *cares* to *care* and *cars* to *car*. An example for Overstemming in our case is our stemmers map both *cheluttuki_nRava_n* (*A man who is riding*) and

chelki_nRava_n (*A man who is going*) to *chel* (*go*) instead of mapping *cheluttuki_nRava_n* to *cheluttu* (*ride)* and *chelki_nRava_n* to *chel*.

If a stemmer removes a nonsuffix from a word, it is called *Mis-stemming*. For example, conflating the words *reply* to *rep* instead of conflating it to *reply* is called *Mis-stemming*. Most of the stemmers do not give importance to *Mis-stemming*. This is because it does not spoil the recall factor in an IR application. Due to the same reason, we evaluate our stemmer using only *Understemming* and *Overstemming*. They are calculated using the following formulas (1) and (2) respectively.

Understemming = (Number of variants understemmed/Total variants)*100% (1)

Overstemming = (Number of variants overstemmed/Total variants)*100% (2)

Evaluating our approach using the above-mentioned corpus containing 765 root variants, we found that 84 and 36 were *understemmed* and *overstemmed* respectively. Hence, The *Understemming* and *overstemming* values are 10.98 % and 4.70 % respectively. Our stemmer effectiveness is calculated using the formula (3).

Stemmer Effectiveness = 100% - [*Overstemming %* + *Understemming %*] (3)

Effectiveness for our approach is 84.32 %. This is considerably a good value. Yet the reason behind achieving a moderate effectiveness value is due to the factors discussed in the Section 4.

7 Conclusion

In this paper, we hypothesize an Iterative Tamil Stemmer. Further, it should be noted that we have evaluated the performance of our stemmer in terms of understemming and overstmming as of now. The proposed stemmers need to be further evaluated with Tamil IR system using factors such as precision, recall etc. Such evaluations will provide the best trade-off between understemming and overstemming that can be obtained by removing or adding a few suffixes in the list. This leads us to believe that our stemmers will prove to be beneficial for Tamil Information Retrieval applications. The limitation with the current version of our stemmers is their ability to handle only suffixes. Investigation on handling prefixes is a part of our future work. It would also be interesting to apply our algorithms to Tamil's Sister languages such as Malayalam, Telugu etc.

References

1. Kraaij, W., Pohlman, R.: Viewing Stemming as Recall Enhancement. In: The Proceedings of the 19th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 40–48 (1996)
2. Germann, U.: Building a Statistical Machine Translation System from Scratch: How Much Bang for the Buck Can We Expect? In: ACL 2001 Workshop on Data-Driven Machine Translation, Toulouse, France (July 7, 2001)

3. Lovins, J.B.: Development of a stemming algorithm. Mechanical Translation and Computational Linguistics 11(1), 22–31 (1968)
4. Porter, M.F.: An Algorithm for Suffix Stripping. Program 14(3), 130–137 (1980)
5. Ramanathan, A., Rao, D.: A lightweight stemmer for Hindi. In: The Proceedings of the 10th Conference of the European Chapter of the Association for Computational Linguistics (EACL) for South Asian Languages Workshop (April 2003)
6. Zahurul Islam, M., Nizam Uddin, M., Khan, M.: A Light Weight Stemmer for Bengali and Its Use in Spelling Checker. In: Proceedings of 1st International Conference on Digital Communications and Computer Applications (DCCA 2007), Irbid, Jordan, pp. 87–93 (2007)
7. Q.-A. Akram, Naseer, A., Hussain, S.: Assas-Band, an affix-exception-list based Urdu stemmer. In: Proceedings of the 7th Workshop on Asian Language Resources (2009)
8. Malayalam Stemmer, http://nlp.au-kbc.org/Malayalam_Stemmer_Final.pdf
9. Kumar, D., Rana, P.: Design and Development of a Stemmer for Punjabi. International Journal of Computer Applications (0975–8887) 11(12) (December 2010)
10. Tamil Noun Flow Chart, http://www.au-kbc.org/research_areas/nlp/projects/morph/NounFlowChart.pdf
11. Tamil Verb Flow Chart, http://www.au-kbc.org/research_areas/nlp/projects/morph/VerbFlowChart.pdf

Appendix: Tamil Transliteration Scheme Used in This Paper

a	aa	i	ii	u	uu	e	ee	ai	o	oo	au	q					
அ	ஆ	இ	ஈ	உ	ஊ	எ	ஏ	ஐ	ஒ	ஓ	ஔ	ஃ					
k	-N	ch	-n	_n	T	N	t	n	p	m	y	r	l	v	J	L	R
க்	ங்	ச்	ஞ்	ன்	ட்	ண்	த்	ந்	ப்	ம்	ய்	ர்	ல்	வ்	ழ்	ள்	ற்

Implementation of Indoor Positioning Using Signal Strength from Infrastructures

Yuh-Chung Lin[1], Chin-Shiuh Shieh[2], Kun-Mu Tu[2,3], and Jeng-Shyang Pan[2]

[1] Department of Management Information Systems, Tajen University,
907 Pingtung, Taiwan
yuhchung@mail.tajen.edu.tw
[2] Department of Electronic Engineering, National Kaohsiung University of Applied Sciences,
807 Kauhsiung, Taiwan
{csshieh,jspan}@cc.kuas.edu.tw,
autoone@ms68.hinet.net

Abstract. Real time location system is not a brand new technology. The most typical approach is using global position system (GPS). However, GPS can only be used outdoors. It is unable to work completely indoors or in an environment with obstacles. Therefore, the development related to indoor position technology is quite important. In the system developed in this study, it makes use of Received Signal Strength Index value of low-power active RFID (Radio Frequency Identification) for movement detection. Besides, it adopts ZigBee wireless transmission technology as reference nodes for positioning detection. Information gathered by reference points is delivered to the server through the Internet. All positioning information is computed by the server. Positioning algorithm uses the average values of signals in its operation. The advantage is to compare many average values with the closing nodes in order to locate the closest position node for the mobile device and reduce the multi-path interference which is caused by other environmental factors. Positioning results can be accessed through the networked computer or mobile device with WiFi functionality. The experimental results are more stable than other positioning algorithms, and the installation of the system is more convenient.

Keywords: Position System, RFID, RSSI, ZigBee.

1 Introduction

Due to the advent of mobility requirement, the real time location positioning demand becomes a significant need in our daily life. The real time location system is not a brand new technology. Recently, the most well-known position system is the Global Position System (GPS) which is based on the signals transmitted by the satellites on the earth orbit. The GPS can locate the longitude and latitude of a place which is determined through triangulation of radio signals transmitted by satellites. In a travel navigation system, GPS users can track their positions depend on the moving velocity and the travel-time of signals transmitted in a line-of-sight propagation condition. In

J.-S. Pan, S.-M. Chen, N.T. Nguyen (Eds.): ACIIDS 2012, Part III, LNAI 7198, pp. 206–215, 2012.

addition to apply to the navigation of car traveling, there are a lot of applications based on the positioning function of GPS. However, the satellite signal can only be received outdoor without any obstacle. It can not be utilized directly for indoor positioning because of the failure of receiving satellite signals. Any obstruction between GPS users and satellites will cause significant errors in location estimation. Therefore, when entering a building or even in a heavy cloudy circumstance, without clear signals of satellites, the GPS will be failed to locate the user's position.

Because of the limitation of GPS, many researchers pay a lot of attentions on the indoor positioning system. There are many ways to implement the indoor positioning system. [1] A simple solution is to implement a network of cameras which is based on the image processing techniques. A network of sensors is also developed as an alternate method for indoor positioning. [2] Recently, Wi-Fi based positioning system becomes popular because it takes advantage of the rapid growth of wireless access point. [3,4] In this paper, we implement an indoor positioning system which makes use of Received Signal Strength Index value of low-power active RFID (Radio Frequency Identification) for movement detection. Besides, it adopts ZigBee wireless transmission technology as reference nodes for positioning detection. Information gathered by reference points is delivered to the server through the Internet. All positioning information is computed by the server.

The rest of paper is organized as follows. Section 2 represents the indoor positioning system implementation. Section 3 shows the experiment results. In section 4, we give a short conlusions.

2 Indoor Positioning System Implementation

2.1 System Architecture

There are different technologies developed for the positioning system, such as GPS, Infra-red, Supersonic, Image processing, RFID and ZigBee positioning. [1] Fig. 1 represents the system architecture in our implementation. The moving object will be equipped with a badge which is a low-power active RFID device as a signal transmitter for the position detection. When the position node receives the packets from the badge, it transmits the packets to the gateway by the ZigBee wireless communication technology. Then the information is delivered to the server through the Internet.

The positioning method is based on the Received Signal Strength Index (RSSI). The system is composed of badges, positioning nodes, servers for location calculation and client PC or smart phone for representing position. The badge and the position node transmit location information to the gateway by wireless communications. The gateway and server are connected by the Internet. Once the location has been calculated, server will send the information to the monitor which is a client PC or handheld devices (ex. smart phone).

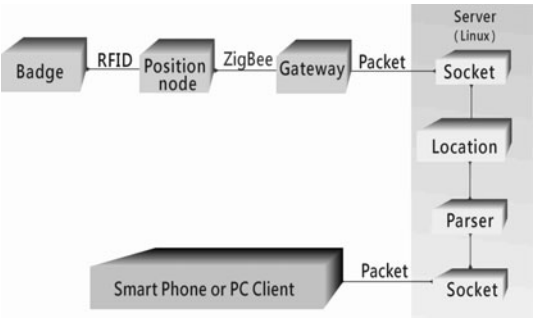


Fig. 1. The architecture of the indoor positioning system

2.2 Transmission Process

There are five steps for the positioning and transmission processes as shown in Fig. 2. At first, the badge will send information to the nearby position nodes. When the data is received, the position node will forward the data to the gateway by a ZigBee connection. Next, the gateway establishes a socket to the server for data transmission. According to the received data, the server will calculate the positioning result via the algorithm that will be described in the following section. Finally, positioning results can be accessed through the networked computer or mobile device with Wi-Fi.

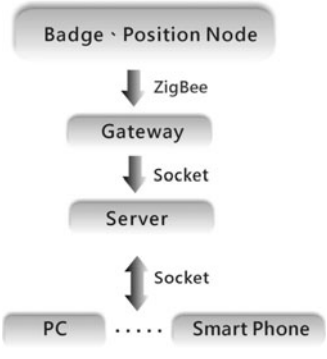


Fig. 2. The positioning and transmission process of the indoor positioning system

2.3 Operations of Server and Client

There are three main components in the server which are the socket program, packet parser and location engine. The socket program is for the connection to the gateway and clients. It opens a TCP socket to be the transmission channel. Once the server receives packets from the gateway, the packets will be delivered to the parser for further processing. The information in the packets will be disassembled to related data structure for the location engine. After a period of time, the location engine will execute a positioning algorithm to process the received information. Afterward the

positioning result will be delivered to the client to show the position of the badge. The processing flow in the server is shown in Fig. 3.

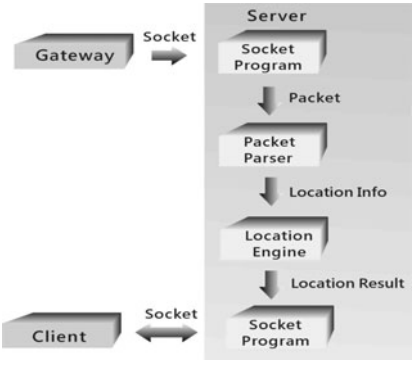


Fig. 3. The processing flow in the server

The client plays the role to display the location result on the end terminal such as PCs or mobile devices. By way of the socket program, the client receives the information from the server. Then it shows the result on the screen. The process is shown in Fig. 4.

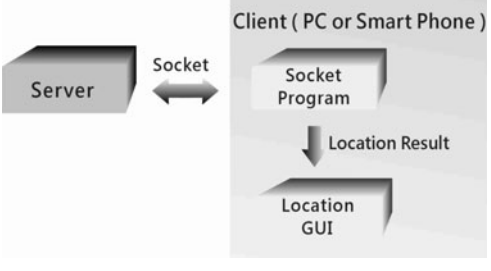


Fig. 4. The processing flow in the client

2.4 Packet Format

Fig. 5 is the data format of the badge and position node. Following is the descriptions of each field. The information sent by the badge includes the MAC address of the badge, the packet ID, the transmitting signal strength, the Link Quality Index (LQI) and RSSI, and the other attached message such as low voltage indication, body temperature, pulse, blood pressure, etc.

The Badge will transmit packets that mentioned above periodically. When the position nodes receive the packet, each position node will append their own ZigBee MAC address to the head of the packet (Fig. 5b) and transmit to the Gateway. Then the gateway transmits all packets that might be received from different position nodes to the server.

In order to measure the signal strength of wireless communication, we utilize the Received Signal Strength Index (RSSI) as a metric to judge the condition of wireless connection. It is widely used in the applications based on the signal strength. The value of RSSI is negative. Therefore, the smaller the RSSI value, the stronger the signal. We also utilize the Link Quality Index (LQI) to represent the link quality of the transmission. The range of LQI is from 0x00 to 0xFF. The higher the LQI value, the stronger the signal.

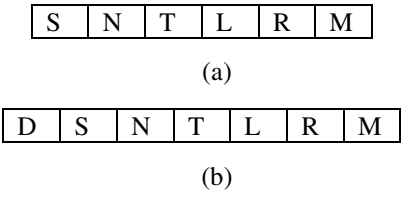


Fig. 5. The packet format of the badge (a) and the position node (b)

D: Position Node's MAC address.
S: Badge's MAC address
N: Packet ID
T: Transmitting Signal Strength of Badge
L: Link Quality Indicator
R: Received Signal Strength Index
M: Other attached messages

The relation of RSSI and transmission distance is shown in the following equation (1).

$$RSSI = -(10n \log 10d + A) \tag{1}$$

where n is the signal propagation constant, d is the transmission distance, and A is the signal strength within 1 meter.

From equation (1), because n and A are constants, RSSI will be decreasing when the distance of two nodes is increasing. Therefore, the distance can be represented as the equation (2).

$$d = 10 \wedge ((|RSSI| - A)/(10 \times n)) \tag{2}$$

Next, equation (3) shows the conversion between RSSI and LQI.

$$RSSI = -(81 - (LQI \times 91)/255) \tag{3}$$

2.5 Positioning Algorithm

In the subsection, we will introduce the positioning algorithm which is used in our implementation. In order to reduce the interference which is caused by other environmental factors, we utilize the average value of RSSIs in the positioning algorithm. The received information will be classified according to the badge's and the position node's MAC address. Based on the pair of (badge, position node), the server will calculate the average RSSI periodically. Then, the smallest average RSSI will be selected and subtract the second small average RSSI. The result will be compared with a threshold. If it is less than the threshold, that means the badge is located between two position nodes. It is hard to determine which position node the badge should belong to. Therefore, the previous position will be maintained. Otherwise, it will compare the received packets' RSSI separately as shown in Fig. 6.

Next, we use an example to explain how the algorithm works. We consider a scenario in which there are three position nodes and one badge. Their IDs are D1=0001, D2=0002, D3=0003 and S=1001. Assume the data received in a period is shown as follows.

D=0001 S=1001 N=FC T=FF L=17 R=40
D=0002 S=1001 N=FC T=FF L=38 R=49
D=0003 S=1001 N=FC T=FF L=3B R=45
D=0001 S=1001 N=FD T=FF L=1E R=40
D=0002 S=1001 N=FD T=FF L=2A R=4C
D=0003 S=1001 N=FD T=FF L=28 R=43
D=0001 S=1001 N=FE T=FF L=17 R=40
D=0002 S=1001 N=FE T=FF L=2C R=4A
D=0003 S=1001 N=FE T=FF L=20 R=45
D=0001 S=1001 N=FF T=FF L=1F R=40
D=0002 S=1001 N=FF T=FF L=31 R=4A
D=0003 S=1001 N=FF T=FF L=37 R=43

Firstly, the server will classify these data into three groups:
D1:
D=0001 S=1001 N=FC T=FF L=17 R=40
D=0001 S=1001 N=FD T=FF L=1E R=40
D=0001 S=1001 N=FE T=FF L=17 R=40
D=0001 S=1001 N=FF T=FF L=1F R=40
D2:
D=0002 S=1001 N=FC T=FF L=38 R=49
D=0002 S=1001 N=FD T=FF L=2A R=4C
D=0002 S=1001 N=FE T=FF L=2C R=4A
D=0002 S=1001 N=FF T=FF L=31 R=4A
D3:
D=0003 S=1001 N=FC T=FF L=3B R=45
D=0003 S=1001 N=FD T=FF L=28 R=43
D=0003 S=1001 N=FE T=FF L=20 R=45
D=0003 S=1001 N=FF T=FF L=37 R=43

Next, the server will calculate the average RSSI of these three groups which are listed as follows.

The average RSSI of D1 is $R1_{avg}$=(40+40+40+40) /4=40
The average RSSI of D2 is $R2_{avg}$= (49+4C+4A+4A)/4= 4A
The average RSSI of D3 is $R3_{avg}$= (45+43+45+43)/ 4= 44

If there is no previous positioning information for the badge, we select the D1 and D3 which average RSSIs are the smallest of the top two. Then the information of the last two packets will be compared as shown in following.

N=FF: D=0001 S=1001 N=FF T=FF L=1F R=40
D=0003 S=1001 N=FF T=FF L=37 R=43
N=FE: D=0001 S=1001 N=FE T=FF L=17 R=40
D=0003 S=1001 N=FE T=FF L=20 R=45

The values of R of D1 are smaller than the one of D3. Then the positioning result will be D1. Otherwise, the next comparison will be proceeded as follows.

N=FF: D=0001 S=1001 N=FF T=FF L=1F R=40
D=0003 S=1001 N=FF T=FF L=37 R=43
N=FD: D=0001 S=1001 N=FD T=FF L=1E R=40
D=0003 S=1001 N=FD T=FF L=28 R=43

If the R of D1 is smaller than the one of D3, D1 will be the positioning result. If not, the final comparison will be proceeded as follows.

N=FE: D=0001 S=1001 N=FE T=FF L=17 R=40
D=0003 S=1001 N=FE T=FF L=20 R=45
N=FD: D=0001 S=1001 N=FD T=FF L=1E R=40
D=0003 S=1001 N=FD T=FF L=28 R=43

If the R of D1 is smaller than the one of D3, D1 will be the positioning result; otherwise the D3 will be the positioning result. The detail of processing flow is shown in Fig. 6.

3 Experiment

The devices that utilized in our experiment are self developed equipments. The wireless communication chip CC2500 [5] of TI is utilized for the transmission of RFID. The badge adopts MSP430F2274 microcontroller of TI to control the CC2500. The communication for ZigBee uses the chip CC2530 [6] of TI which includes the enhanced 8051 microcontroller. We use the MSP430F247 microcontroller of TI as the position node to manipulate the data transmission between CC2500 and CC2530. The Gateway utilizes AX11015 to be the network controller. Fig. 7 is the hardware pictures that are used in our experiment.

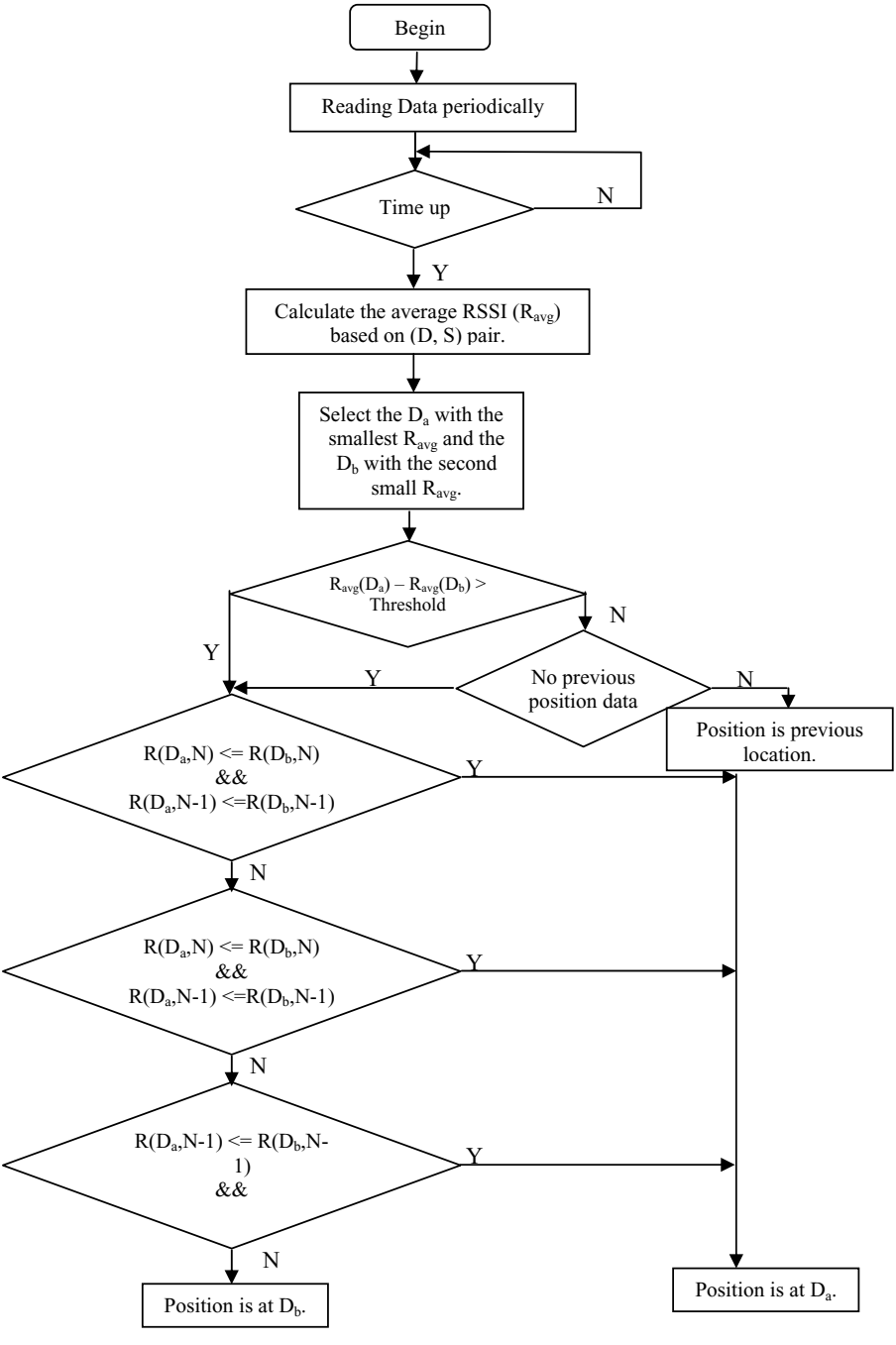


Fig. 6. The flow chart of positioning algorithm

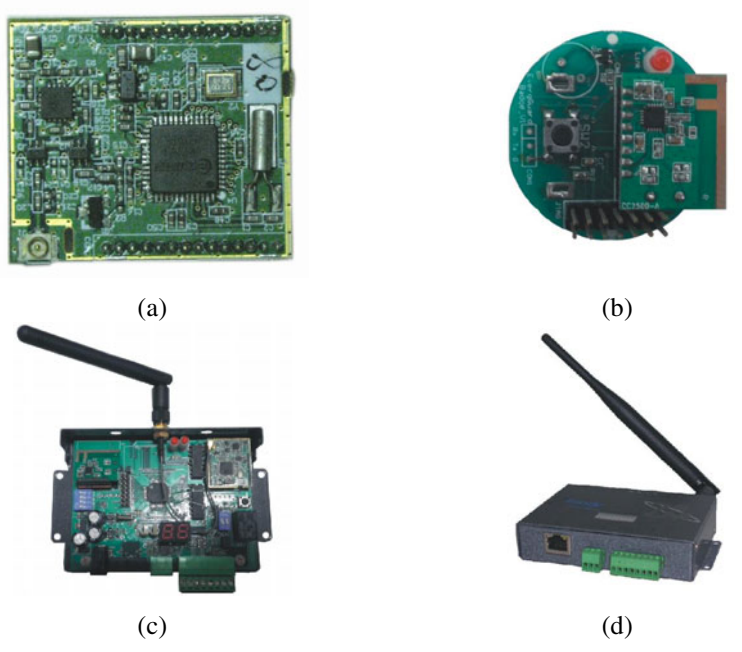


Fig. 7. The pictures of hardware: (a) ZBM_CC2530 Module, (b) Badge Module, (c) Position Node and (d) Gateway

Fig. 8. The experiment environment

Fig. 8 represents the experiment environment. We deploy 10 position nodes and 1 badge. The distance among the position nodes is about 10 meters. The position nodes are the reference points in the positioning system. Therefore, the accuracy of positioning will depend on the density of the position nodes. The signal strength of badge is 0 dbm. The signal will be transmitted once a second. A examiner with the badge walks around all the offices. The positioning algorithm will process data every three seconds.

The examiner carries the badge and monitors the positioning result in the mobile devices (as shown in Fig. 9). When the badge is close to the position node about 3.5

meters, the monitor will show that the badge is approaching to the position node. If the examiner stops about 3 meters away from the position node, within 3 minutes 60 positioning results are obtained. All of them represent the correct position.

Next, the examiner walks around the test environment without stop. The 60 positioning results also are acquired. Some results are not correct. The successful ratio of positioning is about 95%. Then, the examiner increases his walking speed. This will cause the decrement of the successful ratio of positioning which is 60%.

As we can know from the experiment, the moving speed will affect the correctness of positioning. The refresh rate of the positioning algorithm is about 3 seconds. If the transfer time between two position nodes is less than 3 seconds, some data will be viewed as a noise and be ignored. Therefore, the successful ratio of positioning will be reduced. Furthermore, the deployment of more position nodes can raise the accuracy of positioning, but it also reduces the distance among position nodes. This will cause the system is even more sensitive to the moving speed. More frequent transfers among position nodes make the successful ratio of positioning drop dramatically as shown in the experiment. Even though the successful ratio is dropped, it won't affect the accuracy of positioning.

4 Conclusion

In this paper, we implement an indoor positioning system using signal strength from infrastructures. A simple but effective positioning algorithm is proposed. In our experiment, the system can provide a high accurate positioning function. According to our experiment, we know that the density of deployment of the positioning nodes will affect the accuracy of positioning. The experimental results are more stable than other positioning algorithms, and the installation of the system is more convenient.

Acknowledgments. This work is partially supported by the National Science Council, Taiwan, under the grant No. NSC 100-2221-E-151-041-.

References

1. Bejuri, W.M.Y.W., Mohamad, M.M., Sapri, M.: Ubiquitous Positioning: A Taxonomy for Location Determination on Mobile Navigation System. Signal & Image Processing 2, 24–34 (2011)
2. Chang, N., Rashidzadeh, R., Ahmadi M.: Differential Access Points for Indoor Location Estimation. In: IEEE International Conference on Electro/Information Technology, EIT 2009 (2009)
3. Mahtab Hossain, A.K.M., Van, H.N., Jin, Y., Soh, W.-S.: Indoor Localization using Multiple Wireless Technologies. In: IEEE Internatonal Conference on Mobile Adhoc and Sensor Systems, MASS 2007 (2007)
4. Chen, Y., Kobayashi, H.: Signal strength based indoor geolocation. In: Proceedings of the IEEE International Conference on Communications (ICC 2002), New York, NY, USA, vol. 1, pp. 436–439 (April-May 2002)
5. Texas Instruments, CC2500 Datasheet, http://www.ti.com/lit/ds/symlink/cc2500.pdf
6. Texas Instruments, CC2530 Datasheet, http://www.ti.com/lit/ds/symlink/cc2530.pdf

Dual Migration for Improved Efficiency in Cloud Service

Wei Kuang Lai[1], Kai-Ting Yang[1], Yuh-Chung Lin[2], and Chin-Shiuh Shieh[3]

[1] Department of Computer Science and Engineering, National Sun Yat-Sen University, 804 Kaohsiung, Taiwan
wklai@cse.nsysu.edu.tw, terence.kaiting@gmail.com
[2] Department of Management of Information Systems, Tajen University, 907 Pingtung, Taiwan
yuhchung@mail.tajen.edu.tw
[3] Department of Electronic Engineering, National Kaohsiung University of Applied Sciences, 807 Kaohsiung, Taiwan

Abstract. Wireless technologies enable users to retain their Internet access at anywhere and at anytime, without the tangling of wired cables. Users might want to keep enjoying their favor services when they are moving. However, the user mobility would causes longer and longer path to the serving server so that the QoS cannot be guaranteed. In order to maintain better QoS as a user moves, we proposed a novel and efficient cloud service architecture, named dual migration. The dual migration architecture keeps monitor the location of a user and migrates the contents what the user might need onto the closest server for the current location of the user. Therefore, the hop count of the path between a user and the corresponding server is short.

Keywords: Cloud Computing, Service Migration, Server migration, User Mobility, Virtual Machine, Handover, Service Latency, Anycast, Incremental Store.

1 Introduction

Cloud computing is a system architecture for realizing convenient, efficient and of diversity on demand services to users. In the cloud computing architecture, a service provider can share its enormous pool of resources in which storage, computing power, operating systems, software and applications are included with its users so that they can utilize these resources as theirs to achieve their objectives. The service provider can charge management fee for releasing recourses to users.

It has become a promising business model for ISPs, enterprises and users. The fundamental methodology of cloud computing is to separate computation from user devices so that user can enjoy almost all services no matter what equipment they are using. Computing power is provided from a remote server cluster in the cloud. Furthermore, users can store their data in the cloud and make use of applications embedded in cloud system to process their data and retrieve them if needed.

J.-S. Pan, S.-M. Chen, N.T. Nguyen (Eds.): ACIIDS 2012, Part III, LNAI 7198, pp. 216–225, 2012.

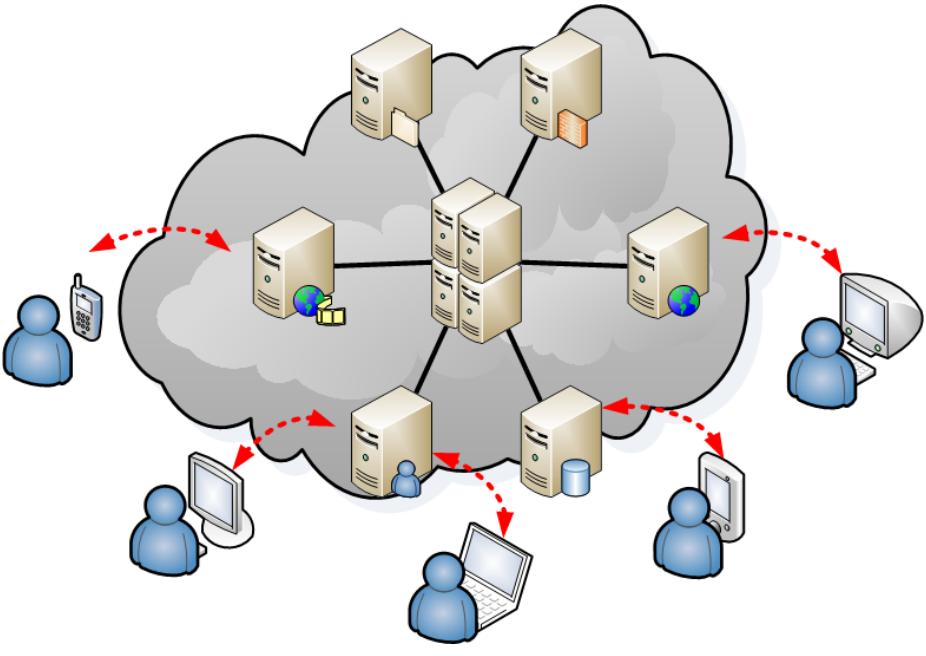


Fig. 1. A Cloud system is composed of many servers which possess powerful computing power and other resourced. Users can access the system and enjoy various services by way of their devices.

Cloud systems provide powerful computation power, scalable storage and other resources to users, so that users do not require mighty hardware as enjoying efficient information services. Users only need to simple, light-weighted devices with input/output and communication capacities to access to cloud systems for their demanded services. As shown in Fig. 1, regardless of their locations, cloud systems provide various services such as high performance computation, multimedia entertainments, word processing, storages and so on to users. After a user demanded a service by his input device, the demand will be transmitted to a cloud system by way of any communication technologies. The cloud system processes the demand and returns a series of results efficiently to the user by taking advantage of the powerful hardware resources and abundant contents in the system. The user then can obtain the result from his output device. Cloud systems let it be possible that users can enjoy complex, various and novel services without being limited by their own equipment.

Although it is scalable, user device-independent and cost saving to separate the storage, computing power and other resources from user devices. It still cause many derivative problems needed to be solved such as service latency, security and so on. It is one of major challenge to reduce the service latency to realize that users can enjoy cloud services as smooth as originating from user devices. It is also an essential factor for the popularity of cloud services.

However, cloud services nowadays are not location-aware. Resources might be far away from users so that the service latencies can't be acceptable. Even cloud system

can finish the user-demanded processes efficiently, it is still a problem that how to return the results to users with acceptable latencies. If the resources are very far away from the locations of users, with considering the heavy traffic along the end-to-end paths, this problem is very difficult to be solved.

Aimed at this problem, we present a novel system prototype for cloud services, named dual migration, which involves the service/content migrations with user migrations (handover) to reduce the service latency of a cloud system and to save the bandwidth of the backbone. The dual migration scheme can make servers and contents are closer to users no matter where users are so that services latencies then can be reduced.

The remainder of this paper is organized as follows. In Section 2, we give a brief introduction to the traditional prototype of cloud system. Representative works related to this study are also reviewed in Section 2. In Section 3, we detail the methodology of the proposed dual migration method. The performance of dual migration is evaluated by simulations, and the results are given in Section 4. Finally, we draw some conclusions of our findings in Section 5.

2 Background Knowledge and Related Works

2.1 Traditional Architecture

In a general cloud service architecture, when a user requests a service, cloud system would assigns one or a cluster of servers to service the user. The assignment might takes into account the server loading, service type, location of users, and so on. If the service is with memory, which means that the current usage must depend on the previous one, such as virtual machine, word processing, and so on, it is essential to load the service status of previous usage before proceeding. The service status is usually stored in the server which previously hosted the user. Therefore, once the user wants to access the same service, the cloud system would assign the previous server to host him again. However, as considering user mobility, the server which is suited to the previous location of the user might not be the best one for the current location of the user. For example, as shown in Fig. 2, if a user accesses a cloud system to request a virtual machine, the cloud system might assign the server which is closest to the location of the user to response the request for efficient transmission. After logout, this server would store the image file of the virtual machine in its storage to guarantee that the user can login the virtual machine again with the same settings. Once the user wants to login the virtual machine again at certain location far away from the server, the connection would be forced to point to the server which stores the image file. Thus, results deduced from user operations would be transmitted to the output device of the user in the form of streams along a very long path. The QoS might be unacceptable. Eventually, after the popularity of cloud services, more streams are transmitted through the backbone, the bandwidth of the backbone would be almost exhausted.

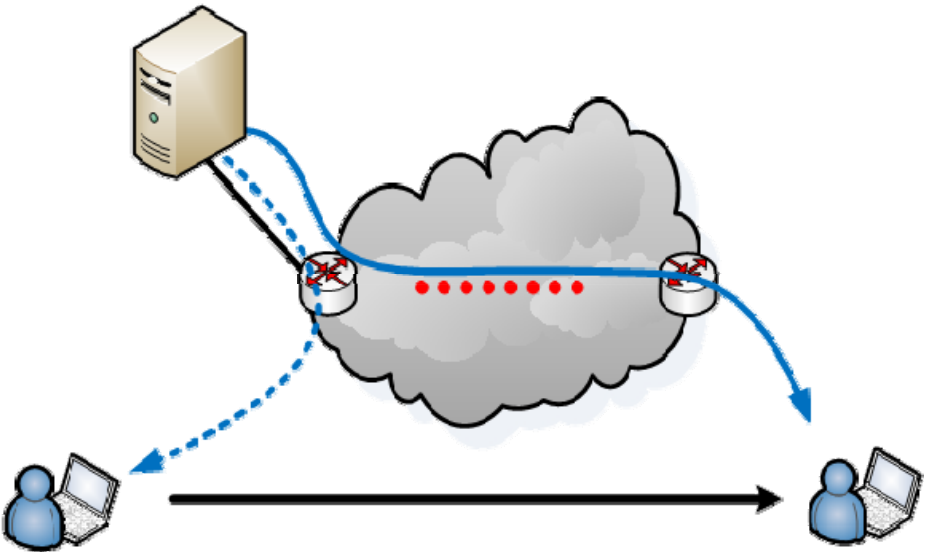


Fig. 2. If the demanded service is with memory, the user is forced to connected to the server which is far away from him. The delay and jitter might be unacceptable

2.2 Related Works

A study for multiple clouds environment had been proposed in [1]. It focuses on the cloud service provisioning from multiple cloud providers. Another research for the interoperability of clouds had been present in [2]. It proposed a five-level model to assess the maturity of cloud to cloud interoperability and gave some suggestions for constructing cloud systems. A resource allocation scheme had been addressed in [3]. The authors of paid their attention to deploy backfill virtual machines for efficient on-demand resource allocation. There had been some researches about service migration [4][5][6], but they did not take the characteristics of cloud computing into their considerations.

3 Proposed Scheme

It is very difficult to maintain better QoS for a long path. In order to provide acceptable QoS no matter where users are, we proposed a novel and efficient cloud service architecture, named dual migration. The proposed DM architecture keeps monitor the location of a user and migrates the contents what the user might need onto the closest server for the current location of the user. Therefore, a user can enjoy services by means of the great capacity of the closest server. The bandwidth of the backbone then can be utilized more efficiently. In other words, we move the data what the user needs to the closest server for the current location of the user, and process it with the aid of the server. Because the distance between a user and the corresponding server is short, the QoS could have more guarantee.

In this paper, there are two difference modes will be discussed. The first one is "hard migration", which means that the content migration is processed after the user logout. This mode is suited for the user who might login the cloud system at any site and never change the location as enjoying the demanded service. The second is "soft migration". In contrast with hard migration, Soft migration means that the content/server migration would come with the user migration (handover) with considering the user mobility. Obviously soft migration has more challenges than hard migration.

3.1 Hard Migration

A user might want to access the same service which is with memory at different locations such as home, office, hotel, etc. It could be an ideal solution for maintaining better QoS that let the server and the content be as close to the user as possible. To realize it, we can direct the user request to the closest server with the aid of anycast [7][8] routing or an intermediate server. As shown in Fig. 3, when a user tries to login the cloud system to access the service with memory again, the request can be directed to the closest server by the router. In anycast methodology, an anycast IP address can be bound with multiple stations. A router with anycast functionality would relay the packet destined to an anycast address to only one station of all stations relative to the address, which is the one whom the router can reach in the least hops. If no router with anycast functionality, the task of directing the request to the nearest server should be performed by an intermediate server of the cloud system. The intermediate server usually directs the request to a server with lower loading, less hops, and more capacity. After the request arrived the most suitable server, in the meantime of accepting the user login, the server contacts with the server which hosted the user last time and downloads the necessary content of the user for the service. According the content downloaded from previous server, the server can host the user with the same setting and the correct progress.

We bring in the "incremental backup" to our proposed architecture with considering the latency of downloading data from previous server. In order to minimize the quantity of the downloading data, each hosted server must store the service progress by using the incremental store method which only stores the difference between the previous usage and the current usage. As shown in Fig. 4, the current hosted server has the same software and libraries as the previous server. Once the necessary data had been downloaded, the current server can process these data and proceeds to serve the user using its own computing power, libraries, software, and other needed resources. Therefore, in the proposed architecture, only the difference between the purely initial service and the status after the latest usage of the user should migrate from previous server to current one. The latency of migrating the necessary data can be considered acceptable.

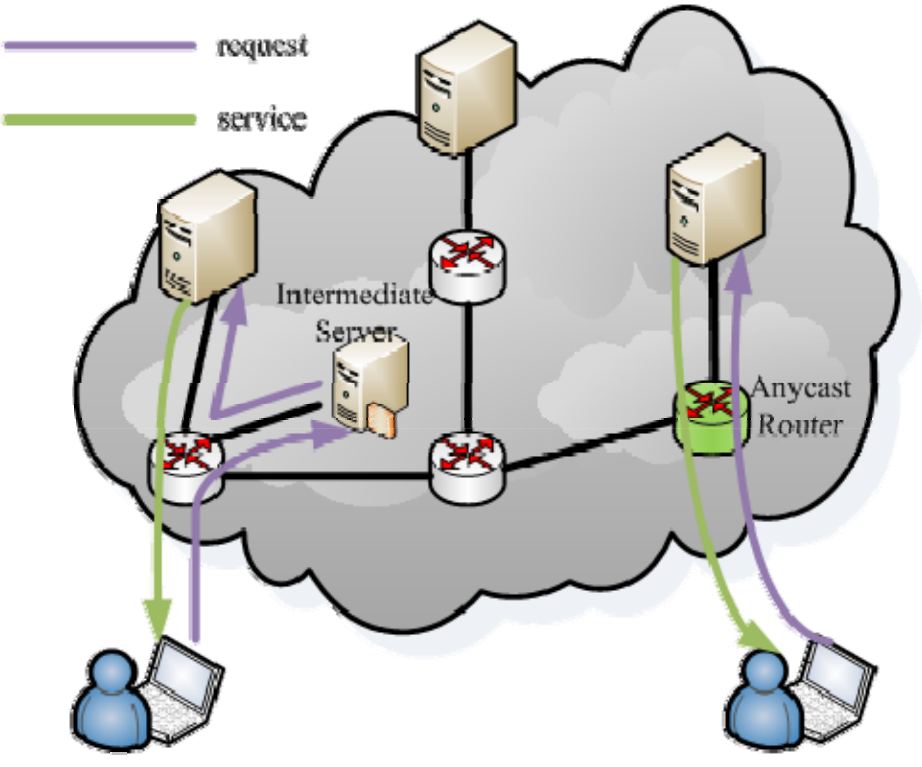


Fig. 3. Users requests can be directed to the nearest server with the aid of an anycast router or an intermediate server

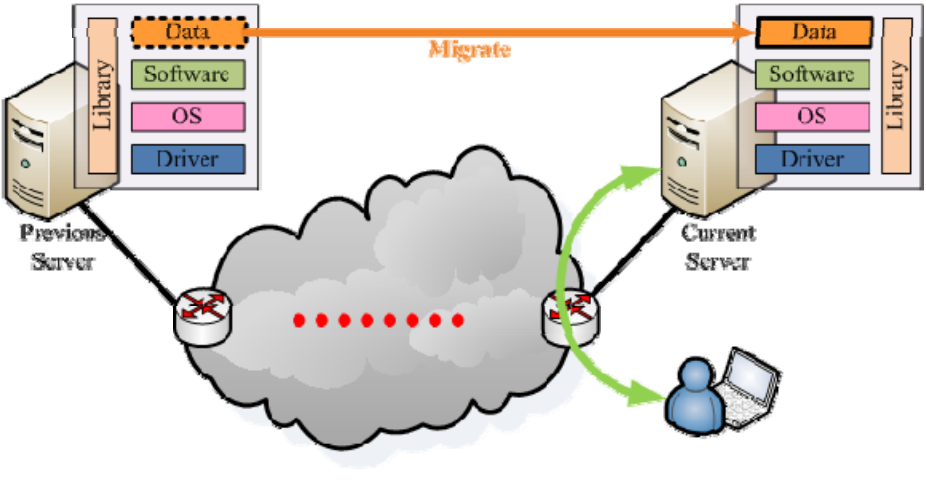


Fig. 4. Only the differential progress would be transferred to the current server. The current server would host the user by its software, library, and so on.

3.2 Soft Migration

Wireless technologies enable users to retain their Internet access at anywhere and at anytime, without the tangling of wired cables. Users might want to keep enjoying their favor services during moving. For this situation, we must consider that how to keep the computing power as close to the user locations as possible no matter where they are going. It is one of the major challenges to maintain the QoS of cloud services with considering the user mobility and the network capacity.

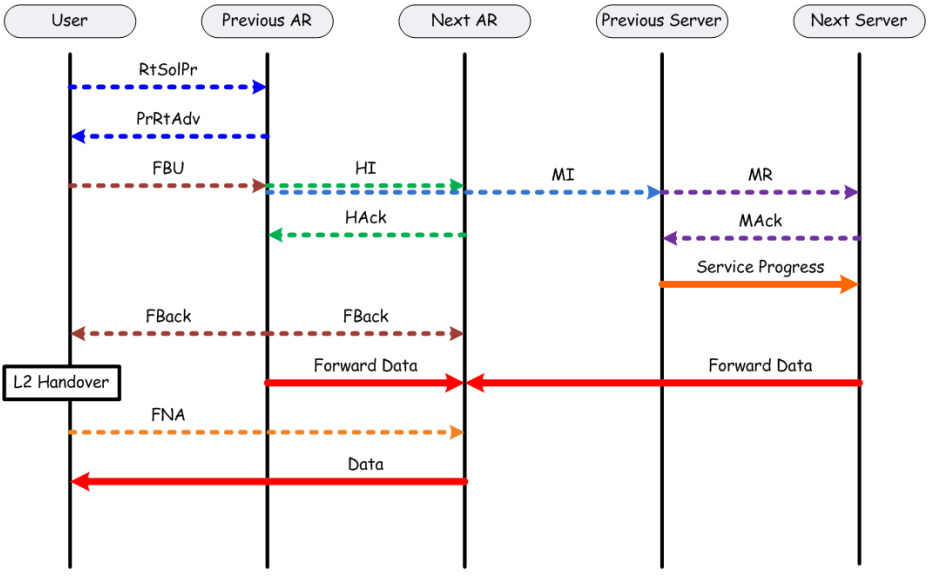


Fig. 5. The flowchart of the dual migration for soft migration mode

Although the methodology is similar as hard migration mentioned previously, the latency constraint is highly stricter than it in hard migration. The reason is that users usually can accept a longer initial delay of starting services such as booting their PCs than the disruption delay when they already enjoy services. For making it possible to reduce the migration delay so that it can be dominated by the originally existed handover delay, the proposed scheme of the soft migration combines the migration procedure with the handover procedure. As shown in Fig. 5, the migration behavior at the server side is also triggered by the layer 2 trigger [9] in Fast Handover, which is specified in RFC 4068[10][11]. In contrast with hard migration, the migration at the server side should come with the handover procedure and should be completed before performing the layer 3 handover which includes re-addressing, binding and so on and is considered of most time-consuming. However the migration latency at the server side is determined by the quantity of data needed to be migrated. It is very difficult to guarantee that the migration can be completed before the layer 3 handover. For this problem, we have two strategies. First, as mentioned previously, we can reduce the quantity of data needed to be migrated as possible. So the incremental store is very

essential for both soft and hard migration. Second, we can let the serving server (previous server) host the user until the migration to the new server is completed. As shown is Fig. 5, before the completion of the server side migration, the serving server keeps serving the user and transferring the service progress and necessary data to the new server simultaneously. Once the new server has all necessary data and knows the latest service progress, the server migration can be regarded as completed; the serving server can terminate hosting the user, and the new server starts to proceed the service.

4 Simulations

We use NS2, a well-known simulator, to validate the performance of our proposed scheme. Fig. 6 shows the scenario in our simulations. At the beginning of our simulation, the user is attached to the router A and receives packets from server 1. Due to the user mobility, the attached router would change as time goes by. In order to reflect the bandwidth contention at the backbone, seven background traffics are added in each router. Each of them generates a 512Kb CBR traffic to the background traffic sink which is attached to the router H. The capacity of each link in our topology is set to 5Mb. The buffer size of each router is set to 100 packets. In the traditional architecture, the user always access the service provided by the server A regardless of the user mobility. On the contrary, the serving server would be changed to the closest one for the user location as the user moves.

Fig. 7 and Fig. 8 show the simulation results in terms of transmission delay and jitter respectively. We can see that the traditional architecture would have higher latency as the increase of the hop count between the server and the user. Similarly, the jitter in the traditional architecture also deteriorate when the hop count increase. Fig. 7 and Fig. 8 also show that our proposed scheme would not be affected by the user mobility. The reason is that our scheme would let the nearest server to provide services to the user. The hop count between the serving server and the user of our proposed scheme is always 2. Therefore both the delay and the jitter are better than those in the traditional architecture.

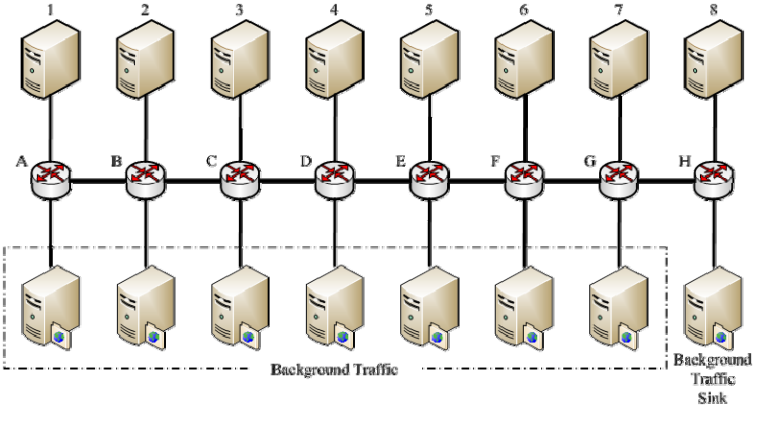


Fig. 6. The Simulation Scenario

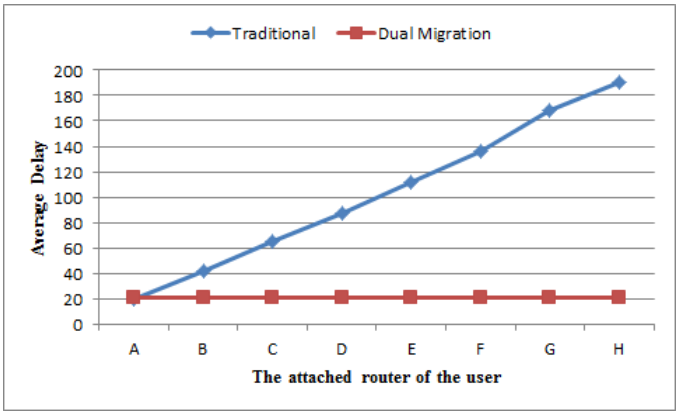


Fig. 7. The results in terms of average delay between the serving server and the user

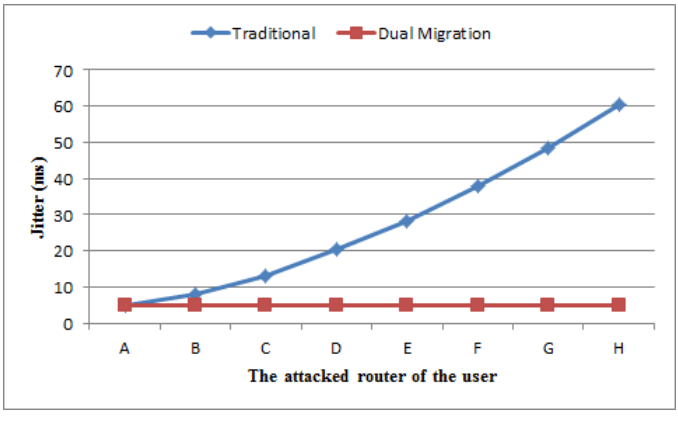


Fig. 8. The results in terms of jitter between the serving server and the user

5 Conclusion

In order to maintain better QoS as a user moves, we proposed a novel and efficient cloud service architecture, named dual migration. The dual migration architecture keeps monitor the location of a user and migrates the contents what the user might need onto the closest server for the current location of the user. Therefore, the hop count of the path between a user and the corresponding server is short. The user can enjoy services by means of the capacity of the closest server. Simulation results show that the dual migration architecture can keep lower transmission delay and jitter no matter where the user is.

References

1. Houidi, I., Mechtri, M., Louati, W., Zeghlache, D.: Cloud Service Delivery across Multiple Cloud Platforms. In: 2011 IEEE International Conference on Services Computing, Washington, USA, pp. 741–742 (2011)
2. Dowell, S., Barreto, A., Michael, J.B., Man-Tak, S.: Cloud to cloud interoperability. In: 6th International Conference on System of Systems Engineering, Albuquerque, USA, pp. 258–263 (2011)
3. Marshall, P., Keahey, K., Freeman, T.: Improving Utilization of Infrastructure Clouds. In: 11th IEEE/ACM International Symposium on Cluster, Cloud and Grid Computing, Newport Beach, USA, pp. 205–214 (2011)
4. Oikonomou, K., Stavrakakis, I.: Scalable service migration in autonomic network environments. IEEE Journal on Selected Areas in Communications 28, 84–94 (2010)
5. Zhao, W., Zhang, H.: Proactive service migration for long-running Byzantine fault-tolerant systems. IET Software 3 (2009) 1751–8806
6. Murakami, K., Haase, O., JaeSheung, S., La Porta, T.F.: Mobility management alternatives for migration to mobile Internet session-based services. IEEE Journal on Selected Areas in Communications 22, 818–833 (2004)
7. Operation of Anycast Services. In: RFC 4786 (2006)
8. Zegura, E.W., Ammar, M.H., Fei, Z., Bhattacharjee, S.: Application-layer anycasting: a server selection architecture and use in a replicated Web service. IEEE/ACM Transactions on Networking 8, 455–466 (2000)
9. Casalicchio, E., Cardellini, V., Tucci, S.: A Layer-2 Trigger to Improve QoS in Content and Session-Oriented Mobile Services. In: 8th ACM/IEEE International Symposium on Modeling, Analysis and Simulation of Wireless and Mobile Systems (2005)
10. Fast Handovers for Mobile IPv6. In: RFC 4068 (2005)
11. Dimopoulou, L., Leoleis, G., Venieris, I.O.: Fast handover support in a WLAN environment: challenges and perspectives. IEEE Network 19, 14–20 (2005)

Integrating Bibliographical Data of Computer Science Publications from Online Digital Libraries

Tin Huynh[1], Hiep Luong[2], and Kiem Hoang[1]

[1] University of Information Technology, Vietnam
[2] University of Arkansas, U.S.A.
{tinhn,kiemhv}@uit.edu.vn, hluong@uark.edu

Abstract. In this paper we proposed and developed a system to integrate the bibliographical data of publications in the computer science domain from various online sources into a unified database based on the focused crawling approach. In order to build this system, there are two phases to carry on. The first phase deals with importing bibliographic data from DBLP (Digital Bibliography and Library Project) into our database. The second phase the system will automatically crawl new publications from online digital libraries such as Microsoft Academic Search, ACM, IEEEXplore, CiteSeer and extract bibliographical information (one kind of publication metadata) to update, enrich the existing database, which have been built at the first phase. This system serves effectively in services relating to academic activities such as searching literatures, ranking publications, ranking experts, ranking conferences or journals, reviewing articles, identifying the research trends, mining the linking of articles, stating of the art for a specified research domain, and other related works base on these bibliographical data.

Keywords: Digital library, data integration, bibliographical data, focused crawler.

1 Introduction

The integration of bibliographical data is considered one of the most important tasks in the area of digital libraries [11]. Searching for the related research publications on the Internet is an important and essential part of work of lecturers, researchers and scientists as well. They can search necessary literature on many different online digital libraries, as well as available indexed databases such as ACM, Springer, IEEE, Elsevier, DBLP, Google Scholar Search, CiteSeer and so on. Just a few years of activities, these digital libraries hold a large collection of books, thesis, journals, especially research publications. Online digital libraries have different methods to collect and to input data into their system. Therefore, the bibliographical databases of these systems will index and save different amount of bibliographical data. Hence, a publication cannot be sometimes found in a specified library, but it can be looked up in others. It is critical to integrate

J.-S. Pan, S.-M. Chen, N.T. Nguyen (Eds.): ACIIDS 2012, Part III, LNAI 7198, pp. 226–235, 2012.

these different data sources into a unified database. This would help to leverage the data search and access for users. Therefore, a data crawling software tool is needed to collect and extract information about articles from different online digital libraries, and then to integrate them into a unified bibliographical database. That tool can go inside many online digital libraries, crawl for new publications and update into the database. In addition, building a system that supports most of services related to academic publishing activities based on social network is a trend of recent researches on exploring digital libraries. Our approach is focusing on using social network analysis and information extracted from this social network, e.g., co-authorship network and citation network, to enhance the integration of digital libraries, especially in the domain of Computer Science publications. In this paper we present a part of our works related to building a software tool to integrate the bibliographical data of publications in the computer science domain from several online digital libraries.

In section 2, we briefly present related research on developing digital libraries, building bibliographical database. Section 3 presents our approach for integrating bibliographical data of publications in computer science, system architecture for metadata extraction from digital library URLs. The assessment of the feasibility of the implemented system will be introduced in section 4. We conclude the paper in section 5 and give some discussions on our system and the future works.

2 Related Work

During the last ten years, projects and software supporting to digital libraries have been developed quickly and have become popular. Greenstone, an open source digital library software, supports librarians in creating collections of science documents [14]. Ian H. Witten et al. also developed a specific tool for Greenstone, i.e., Gatherer, that supports library users the whole process of creating their own collections from digital files. This process contains selecting documents, coming up with metadata set, assigning metadata to each document, building indexes, and putting collections in place for others to use [2]. To integrate bibliographical data from heterogeneous digital libraries, (Schallehn et al., 2001) use XML and related technology for transfering and homogenizing data [11].

According to a survey of research and systems related to the Science Citation Index and Web of Science [9], Eugene Garfield provided the intellectual foundation for the wide variety of citation indexing products [5]. Author in the paper [9], claimed that recent developments in linking of hypertext and web browsers have led to the WoS. As the result, it makes the relationship between citing and cited documents more clearly evident to users. Therefore, bibliographical databases and services offering cited reference indexing, searching need to be developed and built up. Existing systems and services have different data indexing and collecting methods, different data sources and different domain such as Google Scholar, Citeseer, Science Direct, IEEE Xplore, CrossRef and so on listed in [9].

In computer science domain, there are some popular searching and indexing systems for computer science publications such as Google Scholar Search, DBLP, CiteSeer, ACM, SpringerLink, IEEE Xplore, and Microsoft Academic Search.

DBLP (Digital Bibliography and Library Project) provides bibliographic information on major computer science journals and proceedings [8]. DBLP's data are collected by analyzing TOCs (table of contents) of proceedings of conferences, and journals. TOCs files are collected by organizations that have cooperation with the DBLP project and sending to DBLP's authors. Since DBLP is open data, they can be exported to other formats such as CDF, XML and MySQL; developers can download these files from the homepage of DBLP. Based on the data of DBLP, they also developed some searching tools which are introduced in [3], [4].

To check the completeness of data in the DBLP we did a survey to check the existence of publications in DBLP and other online digital libraries. We searched articles on digital libraries with a keyword "information extraction" in the computer science domain, and we got the top one hundred of returned results and checked their existence in the other libraries (Table 1). According to the Table 1 below, DBLP collects most of TOCs files from various conferences, journals organized by ACM, and a small amount of publications of IEEEXplore. Therefore, our system will import existing data of DBLP into our database before integrating bibliographical information of publications from other sources (online digital libraries).

Table 1. Statistical evaluation of the completeness of papers from different digital libraries

100 articles collected from the various online digital libraries	Number of articles found in ACM Digital Library	Number of articles found in IEEE Xplore	Number of articles found in CiteSeer	Number of articles found in DBLP
ACM Digital Library	100	1	41	82
IEEE Xplore	43	100	4	37
CiteSeer	54	24	100	78

Giles et al. present the CiteSeer, an automatic citation indexing system which indexes academic literature in electronic format (e.g. Postscript files on the Web). CiteSeer uses Web search engines (e.g. AltaVista, HotBot, Excite) and heuristics to locate papers (e.g. CiteSeer can search for pages which contain the words "publications", "papers", "postscript", etc.) [6,7]. CiteSeer locates and downloads Postscript files that are available on the internet identified by ".ps", ".ps.Z", or ".ps.gz" extensions, and then it analyzes and extracts bibliographical information from the PDF files downloaded. However, CiteSeer is unable to download papers from online digital libraries that have requested access such as ACM, SpringerLink, IEEE Xplore.

For organizations such as ACM, Springer, IEEE Xplore and other online libraries, they only indexed articles published in conferences, journals that they own. For this reason, some articles exist in a specified library, but not in others (Table 1).

Another framework for scientific digital libraries and search engines with a focus on providing reliable, robust services, i.e, SeerSuite [13], is built by crawling scientific and academic documents from the web. AcaSonet also collects and extracts data of researchers from the Web to build up a social network based on publication lists, and then evaluates publication activities of users in academic community [1]. The Arnetminer system, which aims at extracting and mining academic social network, also integrates publication data into a network from existing digital libraries [12]. It incorporated the extracted researchers' profiles and the extracted publications by using the researcher name as the identifier. However, the data integrating process of this system does not care about domain or subdomain of collected publications. With the aims of collecting, integrating and preparing data for our research on digital libraries search and analytic, our system integrates the bibliographical information of published articles from various online bibliography databases by using topic-focused crawling.

3 Our Approach

Our approach uses a focused crawler (topic spider) to download and analyze the returned web pages that are relevant to a pre-defined topic or set of topics. We launched a search crawling on the list of topics (sub-domains) in the domain of computer science from the ACM computing classification system. Bibliographical information of the crawled articles from various online digital libraries are extracted and imported into database.

3.1 System Architecture

Fig.1 introduces the architecture of our system; the input is a set of keywords (sub-domain) taken from a list of sub-domains in computer science which was captured from the ACM computing classification system. In addition, we also use author names extracted from the DBLP database as an input to the crawling system. Based on the input keywords; the system will automatically create URL queries and submit them to different online digital libraries such as ACM, Springer, IEEE, CiteSeer, MAS (Microsoft Academic Search). The crawling results returned from the digital libraries is a list of URL links of articles that are related to the searching keywords, the bibliographical data extraction module use the pre-defined rules (patterns) to analyze, identify and extract bibliographical information of articles from the returned URLs. The duplicated checking module based on title of returned articles to check if they existed in DBLP, and then the system can store the results obtained if they are new to the database. The import module aims at importing and updating existing data from DBLP into our computer science publications bibliography database.

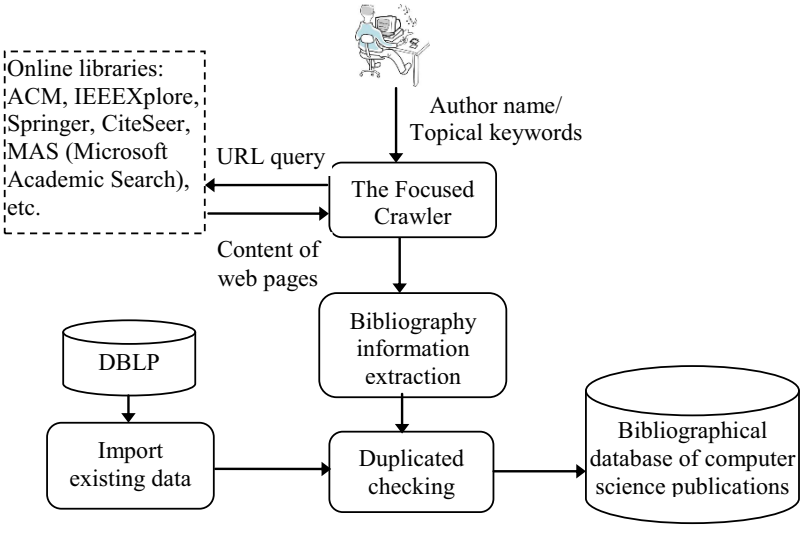


Fig. 1. Architecture of the focused crawling system

3.2 Crawling Papers from the Online Digital Libraries

In order to get relevant papers that match the searching keywords from various online libraries, we submit different queries to these libraries. Then, the system captures links in the returned pages and continues to crawl these new papers links. Currently, our system supports submitting queries to the following digital libraries ACM, IEEEXplore, CiteSeer, MAS (Microsoft Academic Search). In the near future, we are going to implement for other online libraries. The table below presents templates to submit crawling requests to various online libraries.

Table 2. The templates of queries submitted to the relative online libraries

Libraries	Query template	Description
ACM	http://portal.acm.org/results.cfm?query=information%20extraction&dl=ACM&coll=Portal&short=0	The query to ACM with keyword = "information extraction"
IEEE Xplore	http://ieeexplore.ieee.org/search/freesearchresult.jsp?reload=true&queryText=information%20extraction	The query to IEEE with keyword = "information extraction"
Citeseer	http://citeseer.ist.psu.edu/search?q=information%20extraction	The query to Citeseer with keyword = "information extraction"
Microsoft Academic Search (MAS)	http://academic.research.microsoft.com/Search?query=Information%20Extraction&SearchDomain=2	The query to MAS with keyword = "information extraction"

After submitting URL queries to online digital libraries to get down the relative web pages, our tool will analyze these web pages to extract links of relative publications. We implement that similarity for various online digital libraries. Fig. 2 presents an example of extracting links from ACM digital library.

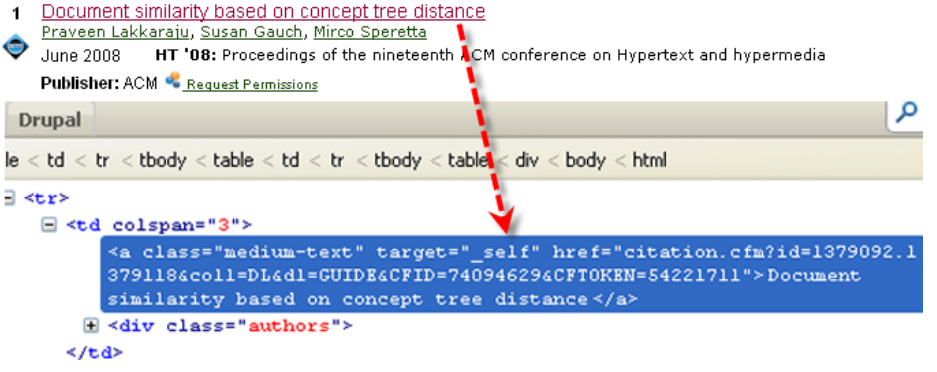


Fig. 2. Analyze the HTML content of webpages to extract links of relative publications

3.3 Bibliography Information Extraction

For each link associated with a relevant publication, we get its BibTeX file by using some predefine rules, patterns. Table 3 presents an example to ID and BibTeX file of a relevant publication from the ACM digital library. BibTeX is a program and file format designed by Oren Patashnik and Leslie Lamport in 1985 [10]. BibTeX uses a style-independent text-based file format for lists of bibliography items, such as articles, books, and theses.

Table 3. Patterns to get id and bibtex file of a relative publication from the ACM

URL Patterns	Description
http://portal.acm.org/exportformats.cfm?id=publicationid	Pattern to get the id of publication base on hyperlink in the HTML content of the returned webpages.
http://portal.acm.org/exportformats.cfm?id=publicationid&expformat=bibtex	Pattern to download a bibtex file of the specified publication from ACM.

Once, we get a BibTeX file of a publication, we use BibTeX parser of JabRef[1] to parse and extract bibliography information of each publication such as author, title, editor, year, pages, numpages, URL, publisher, etc. (Table 4).

[1] http://jabref.sourceforge.net/

Table 4. An example about the structure of BibTeX file of a publication

The structure of BibTeX file
@inproceedings{Dinh:2002:BTC:1118794.1118801, author = {Dinh, Dien}, title = {Building a training corpus for word sense disambiguation in English- to-Vietnamese machine translation}, booktitle = {Proceedings of the 2002 COLING workshop on Machine translation in Asia - Volume 16}, series = {COLING-MTIA '02}, year = {2002}, pages = {1–7}, numpages = {7}, url = {http://dx.doi.org/10.3115/1118794.1118801}, doi = {http://dx.doi.org/10.3115/1118794.1118801}, acmid = {1118801}, publisher = {Association for Computational Linguistics}, address = {Stroudsburg, PA, USA}, }

4 Experiment and Results

In order to evaluate the crawling progress and the ability of enriching bibliographical data of our tool, we compare the crawling results of research articles from three main sources, i.e., ACM, IEEE Xplore and CiteSeer, with the published articles on DBLP that contains 4.873.578 authors and 1.825.233 publications at the experimental time.

First of all we import the DBLP data into the database and then we launch the crawling search with some different keywords. We did testing with keywords such as 'information extraction', 'machine learning' and 'software engineering' on three digital libraries, i.e., ACM, IEEE Xplore and CiteSeer. We take the top 10, top 20, and top 30 publications returned by each digital library. After that we check their existence in existing DBLP to see if they are new publications that should be inserted into the database. The result in Table 5 shows the ability of integrating, enriching and updating new publications from various online digital libraries of our crawling tool.

Specifically, the DBLP database contains most of publications from ACM and CiteSeer libraries, but just a small number of publications from IEEE Xplore digital library. So we can configure suitable values for our crawling tool, for example we can make the schedule to focus most of time on crawling from IEEE Xplore instead of ACM. Integrating bibliography data of publications from various digital libraries in an automatic way is necessary to guaranty the completeness of data for building system that provides academic services relate to research activities.

Table 5. Evaluate updating new publications from various online source

Keyword	Top	IEEE Explorer		ACM		CiteSeer	
		Num. papers existed in DBLP	Correct duplicate checking	Num. papers existed in DBLP	Correct duplicate checking	Num. papers existed in DBLP	Correct duplicate checking
Information Extraction	10	30%	100%	40%	100%	90%	100%
	20	40%	100%	65%	100%	85%	100%
	30	36.66%	100%	73.33%	100%	76.66%	100%
		35.55%	**100%**	**59.44%**	**100%**	**83.89%**	**100%**
Machine Learning	10	40%	100%	70%	100%	90%	100%
	20	35%	100%	80%	100%	85%	100%
	30	46.66%	100%	83.33%	100%	83.33%	100%
		41%	**100%**	**78%**	**100%**	**86%**	**100%**
Software Engineering	10	0%	100%	50%	100%	40%	100%
	20	20%	100%	55%	100%	50%	100%
	30	26.66%	100%	66.66%	100%	63.33%	100%
		16%	**100%**	**57%**	**100%**	**51%**	**100%**

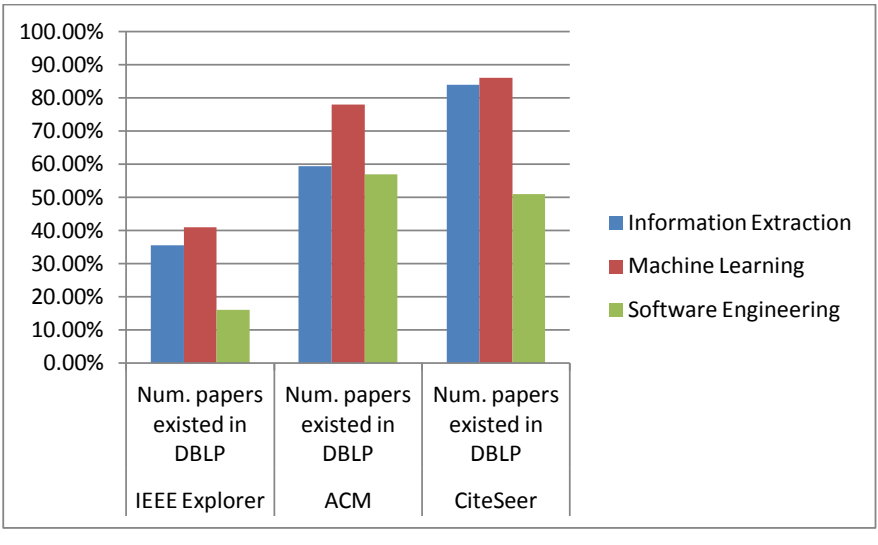


Fig. 3. Number of new articles updated into DBLP

5 Conclusion and Future Work

We have proposed and built a tool to integrate the bibliographic databases of publications in the computer science domain from various online sources into a unified database. Our tool looks for relevant publications by submitting

keywords related to sub-domain of computer science to these digital libraries. We used predefined patterns associated with BibTeX parser to analyze HTML pages and extract bibliographical information of publications.

Unlike ACI tool used in CiteSeer that crawl available PDF files, it may be prohibited by the privacy and access right. Our system does not look for available PDF format file, but it focuses on available bibliographical data of publications instead. Therefore, it can integrate and enrich more bibliographical information of available publications from online digital libraries.

With the collected bibliographical data, we are planning to expand our research related to searching literatures, ranking publications, and experts in the domain, ranking conferences or journals, mining the linking of articles by using social network approach.

References

1. Abbasi, A., Altmann, J.: A social network system for analyzing publication activities of researchers. TEMEP Discussion Papers 201058, Seoul National University; Technology Management, Economics, and Policy Program, TEMEP (2010)
2. Bainbridge, D., Thompson, J., Witten, I.H.: Assembling and enriching digital library collections. In: Proceedings of the 3rd ACM/IEEE-CS Joint Conference on Digital Libraries, JCDL 2003, pp. 323–334. IEEE Computer Society, Washington, DC (2003)
3. Bast, H., Weber, I.: The completesearch engine: Interactive, efficient, and towards ir & db integration. In: CIDR, pp. 88–95 (2007)
4. Diederich, J., Balke, W.T.: FacetedDBLP - Navigational Access for Digital Libraries. Bulletin of IEEE Technical Committee on Digital Libraries 4 (2008) ISSN 1937-7266
5. Garfield, E.: Citation indexes for science: A new dimension in documentation through association of ideas. Science 122(3159), 108–111 (1955)
6. Giles, C.L., Bollacker, K.D., Lawrence, S.: Citeseer: an automatic citation indexing system. In: Proceedings of the Third ACM conference on Digital Libraries, DL 1998, pp. 89–98. ACM, New York (1998)
7. Lawrence, S., Giles, C.L., Bollacker, K.D.: Autonomous citation matching. In: Proceedings of the Third Annual Conference on Autonomous Agents, AGENTS 1999, pp. 392–393. ACM, New York (1999)
8. Ley, M.: The DBLP Computer Science Bibliography: Evolution, Research Issues, Perspectives. In: Laender, A.H.F., Oliveira, A.L. (eds.) SPIRE 2002. LNCS, vol. 2476, pp. 1–10. Springer, Heidelberg (2002)
9. Roth, D.L.: The emergence of competitors to the science citation index and the web of science. Current Science 89(9), 1531–1536 (2005)
10. Sattarzadeh, B., Saffar, Y.G., Asadpur, M.: Making bibliographies using bibtex (2003)
11. Schallehn, E., Endig, M., Sattler, K.U.: Integrating bibliographical data from heterogeneous digital libraries. In: Proceedings of ADBIS-DASFAA Symposium, pp. 161–170 (2000)
12. Tang, J., Zhang, J., Yao, L., Li, J., Zhang, L., Su, Z.: Arnetminer: extraction and mining of academic social networks. In: Proceeding of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD 2008, pp. 990–998. ACM, New York (2008)

13. Teregowda, P.B., Councill, I.G., Fernández, R.J.P., Khabsa, M., Zheng, S., Giles, C.L.: Seersuite: developing a scalable and reliable application framework for building digital libraries by crawling the web. In: Proceedings of the 2010 USENIX Conference on Web Application Development, WebApps 2010, p. 14. USENIX Association, Berkeley (2010)
14. Witten, I.H., Bainbridge, D., Boddie, S.J.: Greenstone: Open-source digital library software with end-user collection building. Online Information Review 25, 288–298 (2001)

Rural Residents' Perceptions and Needs of Telecare in Taiwan

Bi-Kun Chuang [1] and Chung-Hung Tsai[2,*]

[1] Chu Shang Show Chwan Hospital, Taiwan, R.O.C
[2] Department of Health Administration, Tzu Chi College of Technology
880, Sec 2, Chien-Kuo Road, Hualien, Taiwan 97005, R.O.C
tsairob@tccn.edu.tw

Abstract. The purpose of this study was to explore rural residents' perceptions and needs of a telecare system after they have used it. The samples were collected using structured questionnaires with face to face interviews between July 1 and September 30, 2009. Results from this exploratory study show that most elderly people have never heard or touched telecare systems before the study was conducted. However, the general perceptions of such systems include improvement of interacting with medical staffs, safety protection, convenient care, and one needed item of services in daily life. Especially, the mostly risk perception is privacy risk, that is, data confidentiality and individual privacy. Generally, most elderly residents evaluated their telecare experiences and perceptions as being positive. Besides, most elderly resident were willing to use the telecare system without fees. However, they felt risky about confidentiality and privacy toward this technology. To improve trustworthy perception of this novel technology, telecare providers should implement appropriate safeguards to protect patient health information exchanged in a telecare setting. Also, the physicians/nurses should take the time to communicate with the residents, especially in the form of education, about the benefits of technology. To optimize the effectiveness of this promising technique, more research on the relationship between residents' (or patients') perceptions and influences of technology will need to be conducted continually in future.

Keywords: Telecare, risk perceptions, data confidentiality, individual privacy, needs assessment.

1 Introduction

Due to changes in family structure, lowered fertility rate, increased employment rate of women, and weak function of family care [1], it is necessary to combine the resources and support from both the government and health institutions for developing innovative medical service in order to meet rising needs of health care among its elderly people. "Active aging in place" is the main goal of executing healthcare services in a variety of countries [1,2]. Telecare is a method of health care

* Corresponding author.

J.-S. Pan, S.-M. Chen, N.T. Nguyen (Eds.): ACIIDS 2012, Part III, LNAI 7198, pp. 236–246, 2012.

delivery that could address issues of cost and access to care for rural as well as urban underserved patients [3]. Telecare employs information and communication technologies to transfer medical information for the diagnosis and therapy of patients in their place of dwelling. Recently, there are many health care institutions around the world increasingly implementing this innovative technology to improve health care service of elderly people.

Taiwan, has the strength of advanced information technology and popularity of personal computers (66.12%). The Department of Health in Taiwan commissioned telecare projects in 2007 to handle long-term care needs by utilizing innovative information technologies to delivery healthcare to the residents of rural and urban areas. However, most of the research and development of telecare models focused only on medical care [4]. There were relatively few studies on the evaluation of telecare systems in term of general perceptions, risk perceptions, and needs assessment. Understanding perceptions and utility of telecare systems is crucial to optimize design, application, and education strategies that may reduce the burden from caregivers, extend healthy aging in place, and minimize demands on the health care system [5]. How residents (or patients) perceive telecare systems will influence its level of acceptability and consequently its rate of diffusion [3].

Therefore, the purpose of this study is to explore rural residents' perceptions and needs of a telecare system after they have used it. Also, this study will compare the differences in needs of telecare system by gender in these elderly residents.

2 Literature Review

2.1 Development and Application of Telecare

Telecare is defined as "uses modern technology to enable the communication and the transfer of information between the health care provider at the clinical site and the patient at his/her home" [3]. Telecare encompasses a wide range of equipment (e.g., detectors, monitors, alarms, and pendants) and services (e.g., monitoring, call centers and response). Equipment of telecare is offered to support individuals in their home and tailored to meet their specific needs. Telecare services range from a basic community alarm service that is able to react to an emergency to an integrated system that includes detectors or monitors (e.g., falls, fire, and gas) triggering a warning to a response centre. Telecare employs information and communication technologies to transfer medical information for the diagnosis and therapy of patients in their place of dwelling.

The Department of Health of Executive Yuan in Taiwan commissioned telecare projects in 2007 that established three telecare models: community-, home-, and institution-based telecare. This project integrated medical care, medical equipment, information communication technology, and security protection for providing a model of holistic, continued, accessible, and digital healthcare services. Home-based telecare services include three-fold: (1) physiological information retrieval, such as body temperature, heart beat, respiratory rate, blood pressure, blood sugar, and blood oxygen; (2) communication and collaboration of healthcare services, such as urgent

call, transmission of abnormal alarm signals, and notice to revisit; and (3) assistance of health self-management, such as grasping changes of physiological information daily, self-management and follow-up, and early prevention.

Proponents of telecare suggest that it could enable older people to live in a safer and more independent manner. Three generations of telecare systems can be identified. The first generation of telecare systems was technically simple, with no embedded intelligence and entirely reliant on the user activating calls. The second generation systems have all the features of the first generation, but also provide some level of intelligence and automatic detection in limited alert conditions. The third generation systems provide additional support capabilities, such as lifestyle monitoring or reassurance and the introduction of virtual neighborhoods [6].

Two studies investigated perception and needs of patients with chronic illness or home caregivers in telecare services in Taiwan [4,7]. Both studies showed that less than 20% of participants have ever heard or known of telecare. The decision of using home telecare services relied on easy use and operation of telecare equipment, stable and reliable telecare equipment, attitude of staffs, medical assistance of abnormal condition, disease status, needs of daily care, and reimbursement of NHI. Generally, the participants of these two studies have positive attitude towards home telecare. Half the participants agreed to pay fees for using telecare services but only 13.5% of needed telecare caregivers were willing to pay for telecare. More than two thirds of the participants were willing to pay fees for telecare under 1,000 NT dollars. Two studies related to home telecare needs assessment in Canada and United States showed that medical residents have positive impressions on home telecare [3,8]. Also, a low-cost and technologically simple telecare would maximize use of telecare services.

2.2 Needs Assessment and General Perceptions of Telecare

Needs assessment of telecare is a systematic process to acquire an accurate, thorough picture of a system's strengths and weaknesses in telecare, in order to improve it and meet existing and future challenges. The successful introduction of telecare requires the assessment of complex social, political, organizational, and infrastructure characteristics. Multiple factors will determine whether an innovation that is successful or not. One such factor is “readiness”. It is defined as the extent to which a community is ready to attend and succeed in telecare. Four domains of telehealth readiness are patient, practitioner, organization, and public [9].

Defining the need for a telecare service is the first step of a telecare framework followed by planning a service, conducting a needs assessment (a clinical, economic, and technical perspective, developing a healthcare team, marketing, and evaluating the programme [10]. The long-term success of telecare programme and services depends on organizational readiness that is multifaceted concept regarding planning readiness and the workplace readiness. Firstly, planning readiness includes for major themes, like telecare strategic plan, needs assessment and analysis, a business plan, and leadership readiness. Secondly, workplace readiness encompasses human resources readiness (preparing staff, telecare coordinator, and change management

readiness) and structural readiness (technical readiness, policy, access, and communication and participation). A needs analysis is s critical element of organizational readiness. It helps organizations to define their client population and their healthcare problems. It exists to assess how these healthcare needs can be met, and why telecare is the best way of meeting these needs [11].

2.3 Risk Perceptions of Telecare

Telecare uses modern technology to enable the communication and the transfer of information between the health care providers at the clinical site and the patient at his or her home [3]. Because of the potential economic and social benefits of improving access to healthcare, rural areas have been targeted with technology [12]. On the one hand, telecare provides an innovative method of health care delivery that could address the issues of cost and access to care for both rural and underserved patients [13]. On the other hand, the health care sector is also facing many challenges related to the privacy and confidentiality of individual information [14]. Hu, Chau & Sheng [15] argued that perceived risk of telecare might include service efficacy, outcome effectiveness, physician-patient relationships, and patient (information) privacy. They also showed perceived risk was one of significant determinants of telecare technology adoption. Demiris, Doorenbos, & Towle [14] also pointed out that information privacy and confidentiality of ethical consideration were especially important when older adults are the recipients of services. In addition to privacy risk, there also exists performance risk related to the conduct of a virtual visit, including ease of use and effectiveness of telecare equipment [12,13,16]. The present study will investigate two types of risk perceptions of telecare, namely privacy risk and performance risk. The privacy risk includes individual privacy and data confidentiality whereby the performance risk will include ease of use and effectiveness of telecare equipment.

3 Methods

3.1 Sample

We used a self-report questionnaire to empirically examine experiences, general perceptions, risk perceptions, and willingness of using the telecare system. The questionnaire items includes demographic characteristics, health status and behaviors, perception and experience of telecare services, needs assessment of telecare services, and willingness of using telecare services. The survey subjects of questionnaire are the residents who are the end users of a telecare system from Sirliao village, Jhushang township, Nantou County, Taiwan. These end users all had the experience of using the telecare system over one month.

The telecare system was developed and installed by a community hospital in Jhushang township, namely Chu Shang Show Chwan Hospital. The telecare system includes several components: (1) video camera and microphone for online touching case managers of hospital's call center; (2) vital sign monitoring device for detecting

critical health information, such as blood pressure and blood sugar; (3) medical information database for accessing and storing the residents' clinical data.

The criteria for selecting subjects in this study were: (1) the residents of Sirliao village, (2) 65 years or older, (3) had the experience of using the telecare system over one month. There are 277 residents fitted these criteria and selected as the sample. Of the 277 subjects, there were 161 subjects who agreed to participate in the study. The data were collected using structured questionnaires with face to face interviews between July 1 and September 30, 2009.

3.2 Measures

Measures of variables for this study consisted of demographic characteristics; health status and behaviors; general perceptions, risk perceptions, and willingness of using telecare, and telecare needs. Firstly, demographic variables included age, education, marital status, living status, job, financial sources, income per month, and received social welfare and number of financial sources. Secondly, health status and behaviors encompassed chronic disease, number of chronic diseases, type of taking medications, and using medical equipment. Thirdly, perceptions, concerns, and willingness of using telecare services are to measure general perceptions, risk perceptions, and willingness of using telecare. Fourthly, telecare needs were related to safety and security care and health care.

4 Results

Descriptive statistics analysis showed that near half (42%) residents aged 70 – 79 years old and graduated from elementary schools (45%), but females are significantly less educated than males ($p < .001$). Most residents were married (83%) and lived with couple or relatives (87%) in rural areas. Half more residents are farmers (69%) with financial sources from government (73%). Most residents have incomes < 30,000 NT dollars per month (85%) and don't receive social welfare (78%) like home and meals delivered services.

Most elderly people (75%) have one to five types of chronic diseases like hypertension (47%), diabetes (17%), heart disease (19%), joint disease (14%), and chronic obstructive pulmonary disease (COPD, 6%). Surprisingly, females significantly have more chronic diseases ($p = .01$), number of chronic diseases ($p = .03$), joint diseases ($p = .003$), and taken medications ($p = .02$) as compared to males. The majority of elderly residents are taking one to five types of medicine (69%) and using one to four types of medical equipment (74%), such as sphygmomanometer, weight scale, thermometer, and blood sugar meter. Nevertheless, females significantly have more used blood sugar meters than those males ($p = .03$).

Not surprisingly, most of elderly people never have heard (71%) or attended (71%) telecare before as showed in Table 1. Most elderly residents viewed telecare system for improvement of interacting with medical staffs (83%), safety protection (79%), convenient care (79%), and one needed item of services in daily life (78%). The most

important risk perception of using the telecare system is data confidentiality (94%), followed by individual privacy (93%). Most elderly resident (92%) are willing to use the telecare system without fees (88%). Only 11% of residents considered that < 500 NT dollars was acceptable fees to use telecare services regardless safety and security as well as health care. Generally, no significantly differences in experience, perception, concerns, and willingness of using the telecare system between females and males.

Table 1. Experiences, general perceptions, risk perceptions, and willingness of using telecare system

Variable	N (%)		Total N=161 N (%)	Statistical test
	Male N=77 (47.8)	Female N=84 (52.2)		
Experiences of the telecare system				
Had ever heard telecare				$\chi 2(1) = 0.26$,
No	56 (34.8)	58 (36.0)	114 (70.8)	$p = .61$
Yes	21 (13.0)	26 (16.1)	47 (29.2)	
Had ever touched telecare				$\chi 2(1) = 0.02$,
No	16 (31.4)	20 (39.2)	36 (70.6)	$p = .88$
Yes	7 (13.7)	8 (15.7)	15 (29.4)	
General perceptions of the telecare system				
Safety protection				$\chi 2(1) = 1.59$,
Yes	64 (39.8)	63 (39.1)	127 (78.9)	$p = .21$
Unknown	13 (8.1)	21 (13.0)	34 (21.1)	
Convenient care				$\chi 2(1) = 1.59$,
Yes	64 (39..8)	63 (39.1)	127 (78.9)	$p = .21$
Unknown	13 (8.1)	21 (13.0)	34 (21.1)	
Improvement of interacting with medical staffs				$\chi 2(1) = 1.51$, $p = .22$
Yes	67 (41.6)	67 (41.6)	134 (83.2)	
Unknown	10 (6.2)	17 (10.6)	27 (16.8)	
One needed item of services in daily life				$\chi 2(1) = 2.05$, $p = .15$
Yes	64 (39.8)	62 (38.5)	126 (78.3)	
Unknown	13 (8.1)	22 (13.7)	35 (21.7)	
Risk perceptions of the telecare system				
Type of risk perception of telecare				
Individual privacy				$\chi 2(1) = 0.21$,
No	6 (3.7)	5 (3.1)	11 (6.8)	$p = .64$
Yes	71 (44.1)	79 (49.1)	150 (93.2)	
Data confidentiality				
No	3 (1.9)	7 (4.3)	10 (6.2)	
Yes	74 (46.0)	77 (47.8)	151 (93.8)	
Ease of use				
No	73 (45.3)	78 (48.4)	151 (93.8)	
Yes	4 (2.5)	6 (3.7)	10 (6.2)	
Effectiveness				$\chi 2(1) = 0.8$,
No	34 (21.1)	43 (26.7)	77 (47.8)	$p = .37$
Yes	43 (26.7)	41 (25.5)	84 (52.2)	

Table 1. (*continued*)

Variable	N (%)		Total N=161 N (%)	Statistical test
	Male N=77 (47.8)	Female N=84 (52.2)		
Willingness of using the telecare system				
Willing to receive telecare				$\chi 2(1) = 0.5$, p = .48
No	5 (3.1)	8 (5.0)	13 (8.1)	
Yes	72 (44.7)	76 (47.2)	148 (91.9)	
Acceptable fees for safety and security as well as health care per month				
Free	68 (42.2)	74 (46.0)	142 (88.2)	
< 500 NT dollars	8 (5.0)	10 (6.2)	18 (11.2)	
501-999 NT dollars	1 (0.6)	0 (0)	1 (0.6)	

As showed in Table 2, the entire sample ranked the top four safety and security alarm systems of telecare services as urgent call (98.8%), fall detection (95%), dispatching ambulance (75%), and notifying urgent contactor (57%). Likewise, elderly residents ranked the top five telecare services in health care as health counseling (89%), measuring blood pressure (80%), measuring blood sugar (71%), assistance of visiting physicians (63%), and medication reminder (54%). The only significant difference in telecare needs of healthcare by gender is assistance of visiting physicians ($p = .007$).

Table 2. Needs of telecare services

Variable	N (%)		Total N=161 N (%)	Statistical test (2-sided)
	Male N=77 (47.8)	Female N=84 (52.2)		
Safety and security				
Urgent call				
No	1 (0.6)	1 (0.6)	2 (1.2)	
Yes	76 (47.2)	83 (51.6)	159 (98.8)	
Fall detection				
No	2 (1.2)	6 (3.7)	8 (5.0)	
Yes	75 (46.6)	78 (48.4)	153 (95.0)	
Crime events report				
No	74 (46.0)	80 (49.7)	154 (95.7)	
Yes	3 (1.9)	4 (2.5)	7 (4.3)	
Fire system				
No	75 (46.6)	81 (50.3)	456 (96.9)	
Yes	2 (1.2)	3 (1.9)	5 (3.1)	
Looking for missing				
No	72 (44.7)	80 (49.7)	152 (94.4)	
Yes	5 (3.1)	4 (2.5)	9 (5.6)	

Table 2. (*continued*)

Variable	N (%)		Total N=161 N (%)	Statistical test (2-sided)
	Male N=77 (47.8)	Female N=84 (52.2)		
Dispatching ambulance				$\chi^2_{(1)} = 0.17$,
No	18 (11.2)	22 (13.7)	40 (24.8)	$p = .68$
Yes	59 (36.6)	62 (38.5)	121 (75.2)	
Notifying urgent contactor				$\chi^2_{(1)} = 0.1$,
No	34 (21.1)	35 (21.7)	69 (42.9)	$p = .75$
Yes	43 (26.7)	49 (30.4)	92 (57.1)	
Type of safety and security				
1	1 (0.6)	2 (1.2)	3 (1.9)	
2	8 (5.0)	13 (8.1)	21 (13.0)	
3	34 (21.1)	28 (17.4)	62 (38.5)	
4	28 (17.4)	38 (23.6)	66 (41.0)	
5	5 (3.1)	1 (0.6)	6 (3.7)	
7	1 (0.6)	2 (1.2)	3 (1.9)	
Health care				
Measuring blood pressure				$\chi^2_{(1)} = .01$,
No	16 (9.9)	17 (10.6)	33 (20.5)	$p = .93$
Yes	61 (37.9)	67 (41.6)	128 (79.5)	
Measuring sugar				$\chi^2_{(1)} = 0$,
No	22 (13.7)	24 (14.9)	46 (28.6)	$p = 1.00$
Yes	55 (34.2)	60 (37.3)	115 (71.4)	
Measuring heart beats				$\chi^2_{(1)} = 1.5$,
No	53 (32.9)	65 (40.4)	118 (73.3)	$p = .22$
Yes	24 (14.9)	19 (11.8)	43 (26.7)	
Measuring blood oxygen				$\chi^2_{(1)} = 0$,
No	66 (41.0)	72 (44.7)	138 (85.7)	$p = 1.00$
Yes	11 (6.8)	12 (7.5)	23 (14.3)	
Health counseling				$\chi^2_{(1)} = 0.2$,
No	9 (5.6)	8 (5.0)	17 (10.6)	$p = .66$
Yes	68 (42.2)	76 (47.2)	144 (89.4)	
Assistance of visiting physicians				$\chi^2_{(1)} = 7.34$, $p = .007$
No	37 (23.0)	23 (14.3)	60 (37.3)	
Yes	40 (24.8)	61 (37.9)	101 (62.7)	
Medication reminder				$\chi^2_{(1)} = 1.31$, $p = .25$
No	39 (24.2)	35 (21.7)	74 (46.0)	
Yes	38 (23.6)	49 (30.4)	87 (54.0)	
Type of healthcare				
1	1 (0.6)	3 (1.9)	4 (2.5)	
2	10 (6.2)	6 (3.7)	16 (9.9)	
3	28 (17.4)	25 (15.5)	53 (32.9)	
4	20 (12.4)	23 (14.3)	43 (26.7)	
5	6 (3.7)	12 (7.5)	18 (11.2)	
6	2 (1.2)	3 (1.9)	5 (3.1)	
7	10 (6.2)	12 (7.5)	55 (13.7)	

5 Discussion

This study was to explore the health status, health behaviors, experience, general perceptions, risk perceptions, willingness, and needs of an telecare system among elderly residents at a rural community. Generally, the social economic status (SES) of this rural community's residents is relatively low, and most residents have multiple chronic diseases. The low social economic status might affect their adoption of the telecare system. The percent of adoption in telecare was consistent with Chang et al. [7] and Lin et al. [4]. This phenomenon might explain the evidence that the rate of residents regularly using this system is still low.

The unique phenomenon of significantly more chronic diseases, number of chronic diseases, joint diseases, taking medications, and using blood sugar meters in females could not be compared with other studies. This phenomenon might be due to childbearing, diet, labor work and menopause as well as the concern of changes in blood sugar [17].

Most elderly people have positive impression on the telecare system that is similar to Demiris et al. [13], Lin et al. [4], and Scott et al. [8]. However, most elder residents were willing to receive free telecare service. The acceptable fee for telecare services was lower than that in Lin et al. [4], but the percent of willingness to pay for telecare service was similar to Lin et al. [4]. The results might be related to the low social economic status (SES) of this rural community's residents as well as they have fear of using the innovative information technology. Thus, it's necessary to persistently communicate with and educate old residents about cost- benefit analysis and health improvement effects of adopting the telecare system.

With respect to risk perceptions of the telecare system, the highest scoring item was data confidentiality, followed by individual privacy. The results were consistent with Demiris, Speedie, & Finkelstein [13], Turner, Thomas, & Gailiun [12], Hu, Chau, & Sheng [15]. Although residents had positive impression on the innovative technology, they expressed their risk concern about confidentiality and privacy. Trust is a reducer of perceived risk and social uncertainty [18]. People are unwilling to take risks as trust declines. They demand greater protection against possible failures and may be slower to adopt new technology [19]. To improve trustworthy perception of this novel technology, telecare providers should implement appropriate safeguards to protect patient health information exchanged in a telecare setting, including computer security measures and detailed policies controlling the use and disclosure of patient health information [20]. Also, the physicians/nurses should take the time to communicate with the residents, especially in the form of education, about the benefits of technology.

The priorities of telecare needs in safety and security are urgent call, fall detection, dispatching ambulance, and notifying urgent contactor, while the priorities of telecare needs in health care are health counseling, measuring blood pressure, measuring blood sugar, assistance of visiting physician, and medication reminder. This result is consistent with Chang et al. [7]. This population has their unique priorities of telecare needs in safety and security as well as health care that might be related to less educated, low SES, multiple chronic diseases, and inconvenient public transportation.

Especially, females need more assistance of visiting physicians due to no car or no experience in driving cars.

This study has several limitations. Firstly, the content validity and reliability of the self-designed questionnaire have not been tested so that the data may not be highly reliable. Secondly, because of the sample size is limited at a rural community in Taiwan, it's possible that the findings may vary in other community contexts or different countries. Despite these limitations, this study might inspire researchers to develop much more valid and reliable tools to assess the effects of community- and home-based telecare systems among elderly population.

6 Conclusion

With an ageing population, the care of older people and the role of telecare will become increasingly important [6]. Organizational and societal changes, such as cost reduction policies and an aging population, are the main driving forces for the development of telecare systems, especially for elderly patients [21]. In response to importance of telecare technology and limited discussion regarding its adoption by residents, we evaluated the general perceptions, risk perceptions, and needs assessment by conducting a survey study in Taiwan. Generally, our findings supported previous studies. The most elderly residents of this rural community in Taiwan have positive impression on the telecare system and unique priorities in telecare needs. However, owing to their low social economic status, most elder residents wished to receive free telecare service. Furthermore, the residents also expressed their risk perceptions about confidentiality and privacy. These findings seemed to reveal the facts that the telecare system can almost meet most residents' service demands (the accessibility of health care service), but there still exists social psychological issues (risk and trust) to be overcome recently. Hence, successful telecare systems not only provide robust technology quality of information systems, but also delivery reliable service quality of health cares. We believe this study is a useful starting point to explore implement of telecare systems in Taiwan. To optimize the effectiveness of this promising technique, more research on the relationship between residents' (or patients') perceptions and influences of technology will need to be conducted continually in future.

References

1. Ho, T.W., Lai, T.Y.: The Functional Structure Of Taiwan's Pilot Telecare Project. J. of Nursing 55(4), 17–23 (2008)
2. Yang, W.C., Ho, T.W., Huang, T.J., Kung, C.A.: Telecare Pilot Project. ICL Tech. J. 124, 35–38 (2008)
3. Demiris, G., Speedie, S.M., Finkelstein, S.: Change of Patients' Perceptions of TeleHomeCare. Telemed. J. E-Health 7(3), 241–248 (2001)
4. Lin, C.J., Lord, A.Y.Z., Yu, Y.J., Yeh, H.C.: Application Of Telehealth in Case Management. J. of Nursing 56(2), 5–10 (2009)

5. Mahmood, A., Yamamoto, T., Lee, M., Steggell, C.: Perceptions and Use of Gerotechnology: Implications for Aging in Place. J. Hous. Elderly 22, 104–126 (2008)
6. Brownsell, S., Blackburn, S., Aldred, H.: Implementing Telecare: Practical Experiences. Housing, Care and Support 9(2), 6–12 (2006)
7. Chang, T.S., Yeh, M.C., Lou, M.L., Liu, L.F., Hung, L.C.: Comparison Perception Regarding Telehome Care between Home Caregiver and Home Care Nurse. Cheng Ching Medical Journal 3(1), 27–35 (2007)
8. Scott, R.E., Ndumbe, P., Wooton, R.: An E-health Needs Assessment of Medical Residents in Cameroon. J. Telemed. Telecare 11, 78–80 (2005)
9. Jennett, P., Jackson, A., Healy, T., Ho, K., Kazanjian, A., Woollard, R., Haydt, S., Bates, J.: A Study of a Rural Community's Readiness for Telehealth. J. Telemed. Telecare 9, 259–263 (2003)
10. Doolittle, G.C., Spaulding, R.J.: Defining the Needs of a Telemedicine Service. J. Telemed. Telecare 12, 276–284 (2006)
11. Jennett, P., Yeo, M., Pauls, M., Graham, J.: Organizational Readiness for Telemedicine: Implications for Success and Failure. J. Telemed. Telecare 9, 27–30 (2003)
12. Turner, J.W., Thomas, R.J., Gailiun, M.: Consumer Response to Virtual Service Organizations: The Case of Telemedicine. J. Med. Market. 1(4), 309–318 (2001)
13. Demiris, G., Speedle, S.M., Finkelstein, S.: A Questionnaire for the Assessment of Patients' Impressions of the Risks and Benefits of Home Telecare. J. Telemed. Telecare 6, 278–284 (2000)
14. Demiris, G., Doorenbos, A.Z., Towle, C.: Ethical Considerations Regarding the Use of Technology for Older Adults. Res. Gerontol. Nurs. 2(2), 128–136 (2009)
15. Hu, P.J.H., Chau, P.Y.K., Sheng, O.R.L.: Adoption of Telemedicine Technology by Health Care Organizations: An Exploratory Study. J. Org. Comp. Elect. Com. 12(3), 197–221 (2002)
16. Bakken, S., Grullon-Frigueroa, L., Izquierdo, R., Lee, N.J., Morin, P., Palmas, W., Teresi, J., Weinstock, R.S., Shea, S., Starren, J.: Development, Validation, and Use of English and Spanish Versions of the Telemedicine Satisfaction and Usefulness Questionnaire. J. Am. Med. Inform. Assn. 13(6), 660–667 (2006)
17. Deng, Z.H., Su, X.B., Yeh, L.Z.: Understanding and preventing common bone diseases among female menopause. J. of Chinese Orthopaedics and Traumatology 4, 37–43 (2005)
18. Gefen, D., Karahanna, E., Straub, D.W.: Trust and TAM in Online Shopping: An Integrated Model. MIS Quart. 27(1), 52–90 (2003)
19. Hart, P.E., Liu, Z.: Trust in the Preservation of Digital Information. Commun. ACM 46(6), 93–97 (2003)
20. Demiris, G., Doorenbos, A.Z., Towle, C.: Ethical Considerations Regarding the Use of Technology for Older Adults. Res. Gerontol. Nurs. 2(2), 128–136 (2009)
21. Botsis, T., Demiris, G., Pedersen, S., Hartvigsen, G.: Home Telecare Technologies for the Elderly. J. Telemed. Telecare 14(7), 333–337 (2008)

Renewable Energy and Power Management in Smart Transportation*

Junghoon Lee, Hye-Jin Kim, and Gyung-Leen Park**

Dept. of Computer Science and Statistics
690-756, Jeju National University, Jeju City, Republic of Korea
{jhlee,hjkim82,glpark}@jejunu.ac.kr

Abstract. This paper designs a heuristic-based charging scheduler capable of integrating renewable energy for electric vehicles, aiming at reducing power load induced from the large deployment of electric vehicles. Based on the power consumption profile as well as the preemptive charging task model which includes the time constraint on the completion time, a charging schedule is generated as a $M \times N$ allocation table, where M is the number of time slots and N is the number of tasks. Basically, it assigns the task operation to those slots having the smallest power load until the last task allocation, further taking different allocation orders according to slack, operation length, and per-slot power demand. Finally, the peaking task of the peaking slot is iteratively picked to supply renewable energy stored in the battery device. The performance measurement result shows that our scheme can reduce the peak load by up to 37.3 % compared with the *Earliest* allocation scheme for the given amount of available renewable energy.

Keywords: smart grid, charging scheduler, preemptive task model, renewable energy, peak load reduction.

1 Introduction

The future electricity system, called the smart grid, can enhance reliability, efficiency, and security in the energy distribution system, significantly contributing to reducing carbon emissions [1]. The intelligence comes from high-end information technologies such as sophisticated computing algorithms, real-time communication frameworks, and diverse sensor devices. Particularly, with the help of demand response mechanisms, the smart grid can reduce or reshape power load, avoiding the generation of expensive dispatchable energy or new power plant construction [2]. In the mean time, the smart grid will face a serious problem stemmed from the large deployment of electric vehicles, or EVs in short [3]. It is true that EVs reduce greenhouse gases not just by replacing the fossil fuels but

* 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 III, LNAI 7198, pp. 247–255, 2012.

also widely exploiting renewable energy. However, concentrated charging during the specific time interval will put significant stress on the power system.

One of the most promising solutions is the integration of renewable energy such as wind power. A grid unit can run its own electricity generators such as solar cells and wind turbines. For their integration into the grid system, it is necessary to overcome limited dispatchability and unpredictable intermittence. It can be regulated by installing energy storage capacity, as the battery can store electrical energy and be released per requirements [4]. When the power is generated, it will be stored in the battery of charging stations, and used for EV charging afterwards. However, the battery device is not 100% efficient, as some amount of energy is lost in the form of heat due to chemical reactions during charging and discharging processes [5]. Moreover, the battery charge and discharge can't take place at the same time. So, how to manage the battery operation is the most critical issue in integrating renewable energy.

In the intelligent power consumption scenario, demand-side management plays the key role of meeting the customer requirement as well as achieving system goals such as peak load reduction, power cost saving, energy efficiency, and so on [6]. Energy demand is highly likely to be concentrated on a specific time interval, leading to peak load situations which can possibly jeopardize the normal operation of the power generation and distribution systems. For load reshaping, it is necessary to schedule each EV charging operation, which can be modeled as a preemptive task. However, task scheduling is in most cases a complex time-consuming problem which is quite sensitive to the number of tasks. It is difficult to solve by conventional optimization schemes, and the availability of renewable energy makes scheduling more complex. In this regard, this paper designs a power consumption scheduler capable of intelligently integrating renewable energy based on a heuristic and compensatory allocation to peaking slots.

This paper is organized as follows: After issuing the problem in Section 1, Section 2 introduces related work for this paper. Section 3 explains system and task models, respectively, and then designs the scheduling scheme. The performance measurement results are discussed in Section 4. Finally, Section 5 summarizes and concludes this paper with a brief introduction of future work.

2 Background and Wind Energy

To overcome the tremendous computation time imposed by preemptive tasks in generating a power consumption schedule, our previous work has proposed a fast power consumption scheduler which finds a suboptimal solution for the given task set and load profile [7]. For all nonpreemptive task allocations, the scheduler allocates those time slots having the smallest power consumption until the last task allocation to each preemptive task. By modeling a charging operation by a preemptive task, this scheme can work for an EV charging scheduler. It completes in linear computation time, thus removes the exhaustive space search traversal. Taking advantage of such efficiency and speed, not only a variety of task orders can be tried but also their results can be used for the initial population selection

for genetic algorithms. Moreover, the fast allocation mechanism can extend to integrate renewable energy in the charging schedule generation.

In the mean time, transmission distance limitation and instable fluctuation are two major problems of renewable energy development. [8] proposes and implements a parallel planning method for power systems. For a set of partitioned subregions, hour-level and year-level models are built to minimize the power price volatility and the investment cost of transmission. Noticeably, it employs an energy storage system to store surplus power and thus deals with the fluctuation problem of renewable energy. In addition, to improve the economic and environmental sustainability of renewable energy, [9] considers a hybrid system consisting of renewable energy plants and energy storage equipments. Here, its dynamic model is developed, embracing electrolyzer, hydroelectric plant, pumping stations, wind turbines, and fuel cell. Anyway, to cope with the instability of renewable energy, a power storage device is essential.

As for our case, we have been accumulating the wind data from several stations installed at fixed points in Jeju area, which is known to have abundant wind, for the last ten years. Many wind power plants are installed over this area along with wind gauges. Figure 1 plots the monthly wind speed in two locations and we can build a forecast model based on this data using various statistical schemes [10]. The monthly average speed is at least 1 meter per second and almost every day is windy. At this stage, the hourly wind speed and direction data is stored in a relational database table to investigate an appropriate time-series model, which is essential for planning the power generation and storage schedule.

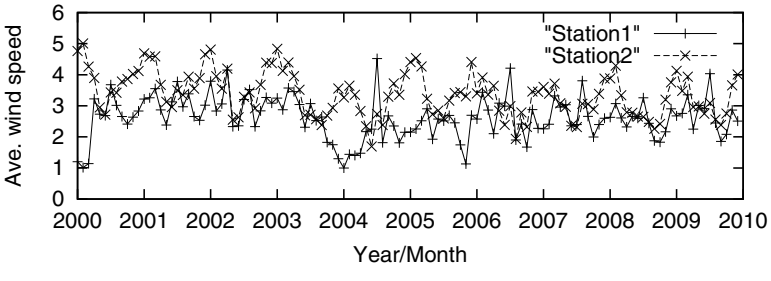


Fig. 1. Monthly wind speed

3 Charging Scheduler Design

3.1 System Model

Figure 2 depicts our system model. A charging request can be issued from EV drivers via either network connection or on-site plug-ins. Each requirement consists of vehicle type, estimated arrival time, desired service completion time (deadline), charging amount, and so on. On every request arrival, the scheduler checks if it can meet the constraints specified by the new request without violating already admitted ones. Receiving the request, the scheduler prepares the

power load profile of the vehicle type from the well-known vehicle information database. The load profile contains the power consumption dynamics along the time axis for specific EV charging [11]. The result is delivered back to the vehicle, and the driver may accept the schedule, attempt a renegotiation, or choose another station.

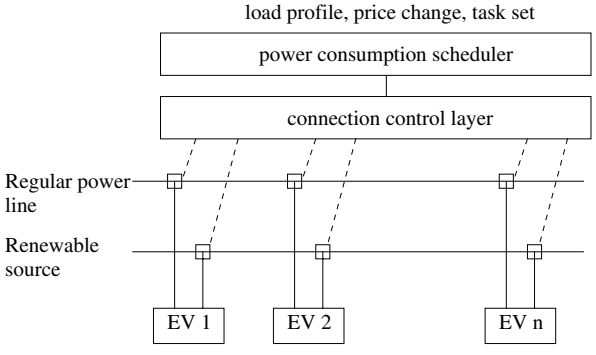


Fig. 2. Charging station model

The scheduler assumes a fixed size time slot for the efficiency and the predictability of a scheduler. Hence, a charging task can begin, suspend, and resume only at the slot boundary. 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. According to the schedule, the connection control layer connects or disconnects EVs to the power source, which can be either the regular power line or the renewable energy. The control layer consists of a lot of controllable on/off switches, while just one source can be connected to an EV for a time slot. In addition, we use a simple linear discharge model for the battery device to focus on the scheduling problem. Namely,

$$V_{disc}^{t} = \alpha \cdot t,$$

where V_{disc}^{t} is the discharged amount until time slot t and α is the discharge coefficient. However, any discharge model can be applied to our scheduling scheme [4].

3.2 Scheduling Scheme

For scheduling, each charging operation can be modeled as a task as in our previous works [7] [12]. 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 consumption profile entry. For more detail, refer to [7] and [13]. 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.

The size of M, namely, the scheduling window, depends on the charging station policy on advance booking.

```
input  : {T_i|(F_i, D_i, A_i, U_i)}   // Task model
         R                              // The amount of renewable energy
output : tab[M][N]                     // Allocation table
var stab[M][N]                         // Temporary allocation table

procedure EvalAlloc
    Δ = 0
    D = {}
    while (R − Δ >= 0)
        select M_j having max(Σ_{i=0}^{M−1} stab[i][j])
        Δ = max(D, M_j) × α
        select M_i having max(stab[i][M_j])
                    where stab[i][M_j] <= R − Δ
        if (found)
            stab[M_i][M_j] = 0
            R− = stab[M_i][M_j]
            D = D ∪ {M_j}
        end if
    end while
    if (max(Σ_{i=0}^{M−1} stab[i][j]) < currentBest)
        replace currentBest and save stab to tab
    end if
end procedure

procedure AllocTab (i)
    if i equals to N − 1
        EvalAllloc ( )             // reaching at a leave
    end if
    for each column from A_i to D_i
        calculate R_i by summing the rows in AllocTab
    select U_i slots having smallest R_i and allocate them to T_i
    map the profile
    AllocTab (i+1)
end procedure
```

Fig. 3. Power scheduling scheme

Figure 3 details our allocation scheme this paper designs to integrate renewable energy. The scheduling procedure takes task set $\{T_i\}$ and the amount of renewable energy, R, to generate the power consumption schedule in $tab[M][N]$. $AllocTab(0)$ is called first and $AllocTab(1)$ will be called next during the space search expansion of $AllocTab(0)$. The procedure recursively continues to T_{N-1}. Then, $EvalAlloc()$ will be called. It selects the peaking time slot, out of which the peaking task is selected. Renewable energy will be assigned to this task at

the time slot. Set D stores those slots which have at least one task assigned to the renewable energy source. Then, the object function is called to evaluate the fitness of the schedule by calculating the peak load. If it is better than the current best, it replaces the current best and the allocation is saved in $tab[M][N]$. After the procedure, $tab[M][N]$ will have the final schedule. Here, Δ denotes the amount of power loss due to time-dependent battery discharge and can be decided by the last slot the renewable energy is used.

Figure 4 shows a simple example of our allocation procedure. There are 4 tasks from T_0 to T_3. Figure 4(a) is the task profile and corresponds to the uncoordinated schedule. α is set to 0 for simplicity to focus on how to allocate renewable energy. However, simple calculation can account for nonzero or more complex discharge coefficients. Figure 4(b) is one of the feasible schedules, that is, every task is allocated within its start time and deadline. Figure 4(c) shows the final schedule when R is 10. At first, the peaking slot is 11, and its peaking task is T_1, so this slot is changed to 0, as it is assigned a renewable energy source. Now, the peaking slot is 6 and its peaking task is 1. The scheduler assigns this slot to the renewable energy. In set D, slots 11, 6, and 5 will be added sequentially, and the power loss is decided by the last slot, namely, 11. In this way, the allocation procedure schedules the operation of power controller. After all, the peak load becomes 5 in Figure 4(c), while it was 11 before the renewable energy is assigned as shown in Figure 4(b).

T0 (2, 4, 15) T1 (5, 5, 14) T2 (1, 6, 10) T3 (5, 6, 19)

	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19
T0	0	0	2	2	3	3	0	0	0	0	0	0	0	0	0	0	0	0	0	0
T1	0	0	0	0	0	4	4	4	3	3	0	0	0	0	0	0	0	0	0	0
T2	0	3	3	3	3	3	3	0	0	0	0	0	0	0	0	0	0	0	0	0
T3	0	0	0	0	0	1	1	2	1	1	1	0	0	0	0	0	0	0	0	0

(a) uncoordinated schedule

	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19
T0	0	0	2	0	0	*2*	0	0	0	3	0	3	0	0	0	0	0	0	0	0
T1	0	0	0	0	0	4	*4*	0	0	0	0	*4*	3	3	0	0	0	0	0	0
T2	0	3	3	3	0	0	3	0	3	0	3	0	0	0	0	0	0	0	0	0
T3	0	0	0	0	0	1	1	0	0	0	0	2	1	1	1	0	0	0	0	0

(b) a feasible schedule

	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19
T0	0	0	2	0	0	0	0	0	0	3	0	3	0	0	0	0	0	0	0	0
T1	0	0	0	0	0	4	0	0	0	0	0	0	3	3	0	0	0	0	0	0
T2	0	3	3	3	0	0	3	0	3	0	3	0	0	0	0	0	0	0	0	0
T3	0	0	0	0	0	1	1	0	0	0	0	2	1	1	1	0	0	0	0	0

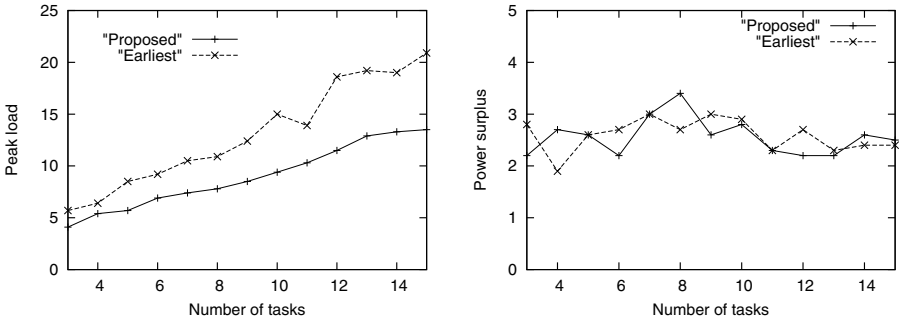
(c) after integrating renewable energy

Fig. 4. Allocation example

4 Performance Measurement

This section implements a prototype of the proposed allocation scheme 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. The experiment sets the schedule length, namely, M, to 20 time units. If a single time unit is 10 min, the total scheduling window will be 3.3 hours, and it is sufficiently large for fast EV charging. For a task, the start time and the operation time are selected randomly between 0 and $M-1$, but it will be discarded and retried 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 has the value of 1 through 5.

Earliest scheduling is selected for the performance comparison as in [12] and [13]. This scheme initiates tasks as soon as they get ready and makes it run without preemption. Even if it adopts no control strategy, it can provide a measure for a comparative assessment for the efficiency of other charging strategies. Actually, it takes too much time to get an optimal solution even for the small number of preemptive tasks by the space search method using the prototype implemented in [7], hence, this section compares the performance just with the *Earilest* allocation scheme. For the *Eairlest* scheme, we also apply the same renewable energy allocation scheme. For each parameter setting, we have generated 50 tasks sets, and their results are averaged.



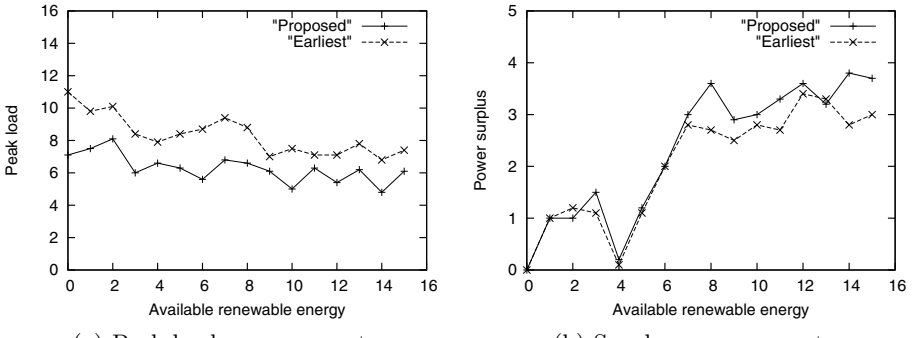
(a) Peak load measurement (b) Surplus measurement

Fig. 5. Effect of the number of tasks

To begin with, our experiment evaluates the performance of the proposed allocation scheme according to the number of tasks. Here, the amount of renewable energy is set to 10. The discharge coefficient is 0.2, thus 4 units will be lost after 20 time slots. First, Figure 5(a) plots the peak load for the number of tasks from 3 to 15. The performance gap gets larger according to the increase in the number of tasks, reaching 37.3 % for 10 tasks and 35.4 % for 15 tasks. Even the same amount of renewable energy is provided with the same compensation strategy, the peak resonance in the noncoordinated allocation is very serious. An efficient task schedule can better benefit from the availability of renewable energy. Next,

the Figure 5(b) shows the power surplus according to the number of tasks. The power surplus is the wasted renewable energy due to time-dependent discharge and uselessness of small pieces. The graph seems not to have a regular pattern, but the proposed scheme wastes smaller amount of renewable energy for the experiment interval.

Next experiment measures the effect of the amount of renewable energy to peak load and power surplus. In this experiment, the number of tasks is set to 5, while the renewable energy amount ranges from 0 to 15. Figure 6(a) compares the peak load values of the *Earliest* scheme and the proposed scheme. Both schemes can reduce the peak load with more renewable energy. The proposed scheme reduces peak load by up to 35.5 %, compared with the *Earliest* scheme, when there is no renewable energy. The gap gets smaller according to the increase on the renewable energy, as the *Earliest* scheme can relieve the peak resonance with more battery power. Hence, the performance gap reaches 17.0 % at the last point. In addition, Figure 6(b) plots the power surplus for the experiment. For more renewable energy, the *Earliest* scheme yields less power waste, and this result is consistent with the result in Figure 6(a).



(a) Peak load measurement (b) Surplus measurement

Fig. 6. Effect of the renewable energy amount

5 Conclusions

In this paper, we have designed a heuristic-based charging scheduler for electric vehicles in the smart grid, for the sake of integrating renewable energy mainly stored in the battery device. To cope with the problem of power load concentration induced by the large deployment of electric vehicles, the scheduler distributes charging operation mainly taking into account the time constraint specified by EVs. Based the preemptive charging task model which consists of the time constraint on the completion time in addition to the arrival time and the operation length, our scheme decides the charging schedule which is represented by $M \times N$ allocation table. Not traversing the vast search space, our basic strategy assigns the task operation to those slots having the smallest power

load until the last task allocation. Then, the peaking task of the peaking slot is iteratively picked to assign renewable energy stored in the battery device. The performance measurement result shows that the proposed scheme can reduce the peak load by up to 37.3 % compared with the *Earliest* allocation scheme for the given amount of available renewable energy.

As future work, we are going to extend the scheduling scheme to augment the price signal change and to save the energy cost. In addition, this scheme will be integrated with a vehicle telematics service according to the available communication infrastructure.

References

1. Gellings, C.W.: The Smart Grid: Enabling Energy Efficiency and Demand Response. CRC Press (2009)
2. Spees, K., Lave, L.: Demand Response and Electricity Market Efficiency. The Electricity Journal, 69–85 (2007)
3. 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)
4. Tremblay, O., Dessaint, L.: Experimental Validation of a Battery Dynamic Model for EV Applications. World Electric Vehicle Journal 3 (2009)
5. Lopes, L., Almeida, P., Silva, A.: Smart Charging Strategies for Electric Vehicles: Enhancing Grid Performance and Maximizing the Use of Variable Renewable Energy Sources. In: Proc. 24th International Battery Hybrid Fuel Cell Electric Vehicle Symposium (2009)
6. Mohsenian-Rad, A., Wong, V., Jatskevich, J., Leon-Garcia, A.: Autonomous Demand-Side Management Based on Game-Theoretic Energy Consumption Scheduling for the Future Smart Grid. IEEE Transactions on Smart Grid 1, 320–331 (2010)
7. 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)
8. Ding, J., Somani, A.: A Long-Term Investment Planning Model for Mixed Energy Infrastructure with Renewable Energy. In: IEEE Technologies Conference, pp. 1–10 (2010)
9. Dagdougui, H., Minciardi, R., Ouammi, A., Robba, M., Sacile, R.: A Dynamic Decision Model for the Real-Time Control of Hybrid Renewable Energy Production Systems. IEEE Systems Journal 4, 323–332 (2010)
10. Palomares-Salas, J., Rosa, J., Ramiro, J., Melgar, J., Aguera, A., Moreno, A.: Comparison of Models for Wind Speed Forecasting. In: Proc. International Conference on Computational Science (2009)
11. Caron, S., Kesidis, G.: Incentive-Based Energy Consumption Scheduling Algorithms for the Smart Grid. In: IEEE SmartGridComm (2010)
12. Lee, J., Park, G., Kim, H., Jeon, H.: Fast Scheduling Policy for Electric Vehicle Charging Stations in Smart Transportation. In: ACM Research in Applied Computation Symposium, pp. 110–112 (2011)
13. Derin, O., Ferrante, A.: Scheduling Energy Consumption with Local Renewable Micro-Generation and Dynamic Electricity Prices. In: First Workshop on Green and Smart Embedded System Technology: Infrastructures, Methods, and Tools (2010)

A Cloud Computing Implementation of XML Indexing Method Using Hadoop*

Wen-Chiao Hsu[1], I-En Liao[2,**], and Hsiao-Chen Shih[3]

[1,2,3] Department of Computer Science and Engineering
National Chung-Hsing University,
250 Kuo Kuang Road, Taichung 402, Taiwan
phd9510@cs.nchu.edu.tw, ieliao@nchu.edu.tw, nic3p1217@gmail.com

Abstract. With the increasing of data at an incredible rate, the development of cloud computing technologies is of critical importance to the advances of researches. The Apache Hadoop has become a widely used open source cloud computing framework that provides a distributed file system for large scale data processing. In this paper, we present a cloud computing implementation of an XML indexing method called NCIM (Node Clustering Indexing Method), which was developed by our research team, for indexing and querying a large number of big XML documents using MapReduce. The experimental results show that NCIM is suitable for cloud computing environment. The throughput of 1200 queries per second for huge amount of queries using a 15-node cluster signifies the potential applications of NCIM to the fast query processing of enormous Internet documents.

Keywords: Hadoop, Cloud Computing, XML Indexing, XML query, Node Clustering Indexing Method.

1 Introduction

XML (eXtensible Markup Language) is widely used as the markup language for the web documents. The flexible nature of XML enables it to represent many kinds of data. However, the representation of XML is not efficient in terms of query processing. A number of indexing approaches for XML documents are proposed to accelerate query processing. Most of these works provide mechanisms to construct indexes and methods for query evaluation that deal with one or small amount of documents in a centralized fashion. In the real world, an XML database may contain a large number of XML documents which require the existing XML indexing methods to be scalable for high performance.

The concept of the "cloud computing" has been received considerable attention because it provides a solution to the increasing data demands and offers a shared,

* This research was partially supported by National Science Council, Taiwan, under contract no. NSC100-2221-E-005-070.

** Corresponding author.

J.-S. Pan, S.-M. Chen, N.T. Nguyen (Eds.): ACIIDS 2012, Part III, LNAI 7198, pp. 256–265, 2012.

distributed computing infrastructure [2]. With the increasing popularity of cloud computing, Apache Hadoop has become a widely used open source cloud computing framework that provides a distributed file system for large scale data processing. When a low-cost, powerful, and easily accessible parallel computational platform is available, it is important to better understand how it can solve a given problem [3].

Although there are many published papers on the subject of XML indexing and querying methods, most of them are confined to small data samples running in the centralized system. As cloud computing becomes popular, the issues of parallel XML parsing have been discussed recently. However, to the best of our knowledge, there is very little work that addresses the problem of indexing as well as querying XML documents on large distributed environments. Exploring whether the existing XML indexing methods can be scaled out is an important issue due to the enormous XML documents in the Web.

In our previous work [1], we presented an indexing method called NCIM (Node Clustering Indexing Method) which compresses XML documents effectively and supports complex queries efficiently. In this paper, we use Hadoop framework to present a mechanism for distributed construction and storage of indexes as well as distributed query processing for a large number of big XML documents on the basis of NCIM.

The contributions of our work are as follows. We modify the NCIM (Node Clustering Indexing Method) and design a system for indexing and querying a large number of XML documents by using the Hadoop cloud computing framework. We also consider two job processing modes, streaming query vs. batched query, for query evaluation in our experiments. The results show that the batched query processing will have much better throughput.

The rest of this paper is organized as follows. In the next section, we review related work. Section 3 describes preliminaries on Hadoop. Section 4 presents the proposed system that builds indexes for XML datasets and answers massive queries simultaneously. Experimental results are discussed in Section 5. Finally, Section 6 concludes the paper.

2 Related Work

Many index methods and query evaluation algorithms have been proposed in the literature. The most widely used approaches are structural summary and structural join. The structural summary indexing methods merge the same sub-structures in an XML document and form a smaller tree structure, which is used as the index of the XML document. Thus, instead of matching an input query against the XML document itself, the summarized index tree is used. The DataGuide [4] is a typical model. A strong DataGuide holds all the P-C (Parent-Child) edges in an XML file. Each node in a DataGuide has an extent for the corresponding nodes in the original XML document. Therefore, the P-C (Parent-Child) and A-D (Ancestor-Descendant) relationships can be evaluated using strong DataGuide directly. However, DataGuide is not feasible for twig queries, since the structure of the summarized index is not the same as the original XML document.

Structural Join [5] is one of the first proposed methods to process twig pattern matching. A twig query is decomposed into several binary P-C or A-D relationships. Each binary sub-query is separately evaluated and its intermediate result is produced. The final result is formed by merging these intermediate results in the second phase. This method generates a huge amount of intermediate results that may not be part of the final results. In addition, the phase of merge is expensive. Various follow-up techniques have been proposed to filter out useless partial solutions and avoid the expensive merging phase [6, 7, 8].

The NCIM [1] method labels each element node of an XML data tree with 3-tuple (level, $n^{\vdash}$, $n^{\dashv}$) for non-leaf node and a 2-tuple (level, $n^{\vdash}$) for leaf node, where "level" is the depth of the node n with the root as level 1, "$n^{\vdash}$" (start number) is the serial number of node n derived from a depth-first traversal of the data tree (the root node is assigned 1 also), and "$n^{\dashv}$" (end number) is the serial number after visiting all child nodes of n. The information is clustered with same (tag, level) pair and stores them in four hash-based tables, two for node indexes and two for level indexes. The advantage of using hash tables is to gain fast accesses on the needed data. NCIM can deal with single-path query as well as more complex query patterns. The experimental results show that NCIM can compress XML documents with high compression rate and low index construction time. There have been many indexing methods proposed in the literature for XML query processing. However, very few are known to scale out for a large number of big XML documents.

In recent years, parallel XML parsing and filtering have been discussed for processing streaming XML data in scientific applications. The results of parsing XML can be DOM-style or SAX-style. The parallel DOM-style parsing constructs a tree data structure in memory to represent the document [9, 10, 11]. The load-balancing scheme is widely applied that assigns work to each core as the XML document was being parsed. The parallel SAX-style parsing visits XML document in depth-first traversal. It is much more suitable when XML documents are streaming. In Pan et al. [12], they present algorithms on how to parallelize the parsing computations prior to issuing the SAX callbacks (representing the events). Although some of parallel XML parsing techniques have been proposed, indexing and querying XML documents on large distributed environments remains a challenging issue.

3 Preliminaries on Hadoop

The Apache Hadoop software library, inspired by Google Map-Reduce and Google File System, is a framework that allows for the distributed processing of large data sets across clusters of computers using a simple programming model [13]. Hadoop consists of two main services: high-performance parallel data processing using a technique called MapReduce and reliable data storage using the Hadoop Distributed File System (HDFS). Since Hadoop is well suited to process large data sets, the proposed system uses Hadoop as the cloud computing framework.

The MapReduce, illustrated in Fig. 1, has two computation phases, map and reduce [14]. In the map phase, an input is split into independent chunks which are distributed

to the map tasks. The mappers implement compute-intensive tasks in a completely parallel manner. The output of the map phase is of the form <*key, value*> pairs. The framework sorts the outputs of the mappers, which are then passed to the second phase, the reduce phase. The reducers then partition, process and sort the <*key, value*> pairs received from the Map phase according to the *key* value and make the final output.

The HDFS is a distributed file system designed to store and process large (terabytes) data sets. HDFS is highly fault-tolerant and is designed to be deployed on low-cost hardware. HDFS provides high throughput access to application data and is suitable for applications that have large data sets [15]. HDFS has a master/slave architecture that consists of a single NameNode and a number of DataNodes. The NameNode, the master of the HDFS, maintains the critical data structures of the entire file system. The DataNodes, usually one per node in the cluster, manage storage attached to the nodes that they run on. Internally, a file is split into one or more blocks that are stored in a set of DataNodes with replication. The NameNode executes file system namespace operations like opening, closing, and renaming files and directories. It also determines the mapping of blocks to DataNodes. The DataNodes are responsible for serving read and write requests from the file system's clients. The DataNodes also perform block creation, deletion, and replication upon instruction from the NameNode [15].

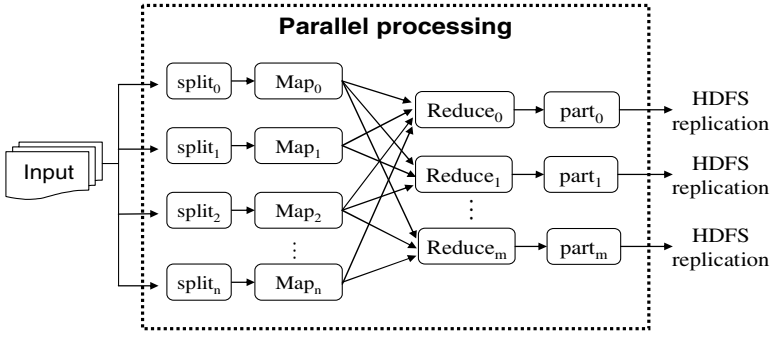


Fig. 1. The processing of MapReduce

Hadoop MapReduce framework runs on top of HDFS. Typically the compute nodes (the Map/Reduce framework) and the storage nodes (the HDFS) are running on the same set of nodes [14]. Thus data processing is co-located with data storage. A small Hadoop cluster will include a single master and multiple slaves. Fig. 2 [16] shows a hadoop system with multi-core cluster. The master for the MapReduce implementation is called the "JobTracker", which keeps track of the state of MapReduce jobs and the workers are called "TaskTrackers", which keep track of tasks within a job. The job tracker assigns jobs to available task tracker nodes in the cluster as close to the data as possible. If a task tracker fails or times out, that part of the job is rescheduled by job tracker.

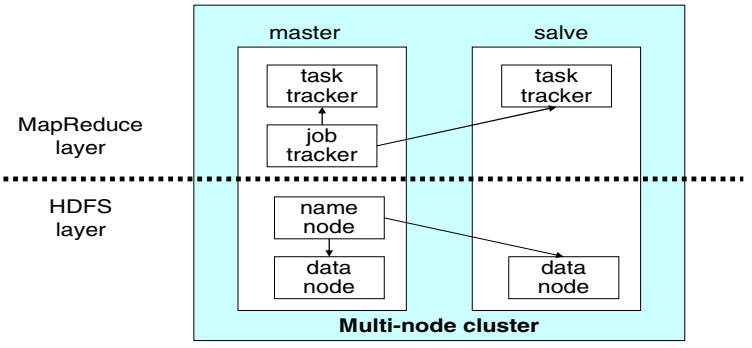


Fig. 2. The Hadoop system overview [16]

Since the release of Hodoop system, more and more researches use Hadoop as a framework to develop applications on large-scale data sets. For example, Zhang, et al. [17] describe a case study that uses the Hadoop framework to process sequences of microscope images of live cells. Dutta, et al. [2] address the problem of time series data storage on large distributed environments with a case-study of electroencephalogram using Hadoop.

4 The Proposed System

Consider a database that contains a large number of XML documents with different structures and sizes. It is a challenging task to retrieve required information from such huge amount of documents. In this paper, we develop a Hadoop-based XML query processing system for indexing and querying large number of XML documents. The proposed system consists of two subsystems. The first one is the preprocessor that parses XML documents and builds indexes. The second subsystem is the query processor that accepts queries from users and does query evaluation with the help of indexes.

4.1 Index Construction

The indexing method used in the proposed system is NCIM with some modification. We choose NCIM because the experimental results show that NCIM can compress XML documents effectively and support complex queries efficiently. The original NCIM method constructs and stores the indexes, which consist of four hashed-based tables, in main memory to support fast accesses. However, this implementation is not suitable for big data size of XML documents. Therefore, we write the indexes into files in the proposed system. The Non-leaf node index and the leaf node index are stored based on the hash keys. That is, data in each linked list to which a hash entry points is written into a file named using the corresponding hash key. Using this storage strategy, we need only to load required data while a query is processed. There is another modification in the proposed implementation of NCIM. Only text contents

which are less than 20 characters in the leaf node index are stored in the file for saving space. However, this restriction will be lifted in the future implementation for supporting wildcard characters in the queries.

In this phase, the input (see Fig. 1) is a sequence of files and each file represents an XML document. We refer to each file as a split, according to the Hadoop terminology, and feed splits to the map tasks. The splits are then processed in parallel. The SAX parser is used to parse the input XML document and the modified NCIM method is used to construct indexes. These two utilities reside in the Map function. Each Map function produces a list of <*key, value*> pairs, where *key* is a (*tag, level*) pair and the *value* is the corresponding *label* of a node. After that, the MapReduce framework collects all pairs with the same key from all lists and groups them together. The Reduce function is then applied in parallel to each group, which in turn produces a collection of values in the same *key,* and then the results are written into files. The output files may reside in different data nodes depending on HDFS. The flow of index construction is shown in Fig. 3.

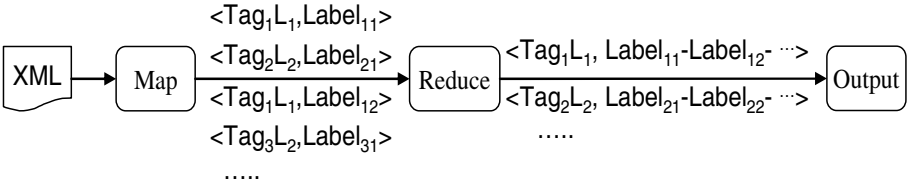


Fig. 3. The flow of index construction

4.2 Query Evaluation

After building XML indexes in the cloud servers, users may start sending queries to the cloud servers. The cloud servers may receive tens of thousands of queries per second and must be designed to respond in a very fast way. Fig.4 shows the overview of the proposed XML-Cloud service system. The cloud frontend receives user queries and submits them to the JobTracker. The JobTracker decides how many map tasks are required and distributes these tasks to the chosen TaskTracker nodes. Because the indexes are stored in HDFS, the JobTracker schedules tasks on the nodes where data is present or nearby. The TaskTracker loads indexes from HDFS and performs query evaluation. In this case, no reducer tasks are needed.

The query evaluation in our system is based on the algorithm of NCIM. The difference is that NCIM keeps indexes in main memory and the proposed system saves indexes in HDFS. Loading corresponding indexes are necessary before doing query evaluation in our system.

In the proposed system, we design two query processing modes. One is called streaming mode (Mode I). The second one is called batch mode (Mode II). In streaming mode, we treat user queries as a stream, and then each query is treated as a split and assigned to a map task. However, this may require a large number of I/O because each query will load corresponding part of indexes for evaluation. Two queries over the same documents may be assigned to different machines. The

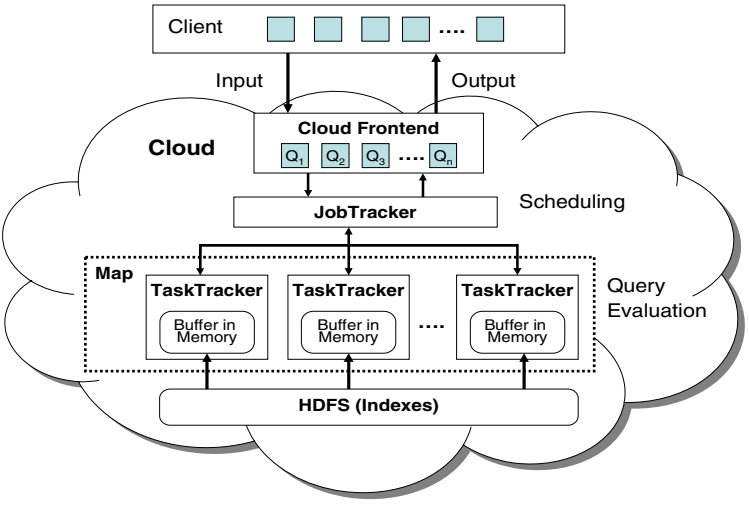


Fig. 4. The overview of XML-Cloud service system

common parts of indexes are loaded and released in different nodes. This will impose high cost for servers and result in poor performance.

In the batch mode, user queries are collected for a time period, e.g., one second, in the cloud frontend, and then classified into groups according to some similar characteristics. A query group is then treated as a split and assigned to a machine that loads common parts of indexes once and releases them after all queries in a group are finished. There may be a delay between query being entered into the system and the query being processed. However, the throughput of the system increases substantially when there are many user queries at the same time.

5 Experimental Results

In the experiments, we use a homogeneous cluster, which is composed of 15 slave nodes on Linux X86-64 with 4 CPUs and 8GB DRAM[1]. The Hadoop 0.20.1 is installed to run the experiments. Because it is a small cluster, the test datasets are not big comparing to real world datasets. The maximum size of datasets is about 2.5GB, which contains 50 XML files with different sizes. The maximum and minimum size of XML files are 75MB and 32MB, respectively. All essential functionalities are put inside the cloud. There is also a simple user interface at the client side for submitting XML documents and queries.

5.1 Performance of Index Construction

In the index construction phase, a set of XML files are parsed, and indexes are produced, which are then stored as HDFS files. In order to evaluate the performance

[1] Thanks to the National Center for High-Performance Computing, Taiwan, for providing the Hadoop computing cluster.

of the system under different data size, we form 5 datasets of size 0.5GB, 0.9GB, 1.4GB, 1.9GB, and 2.5GB with 10, 20, 30, 40, and 50 files, respectively. Fig. 5 shows the execution time for one file per task. The execution time includes loading XML files, executing MapReduce program, and saving index files to HDFS. It can be seen that the execution time is not increased linearly in terms of the number of files. The reason is because there are 15 slave nodes in the cloud, and the tasks are distributed unevenly when the number of input files is not the multiple of 15. We also observed that the Hadoop spent most of times in reduce tasks, where the HDFS creates multiple replicas of data blocks and distributes them on compute nodes. The *replication factor* is set to 3 for all tests.

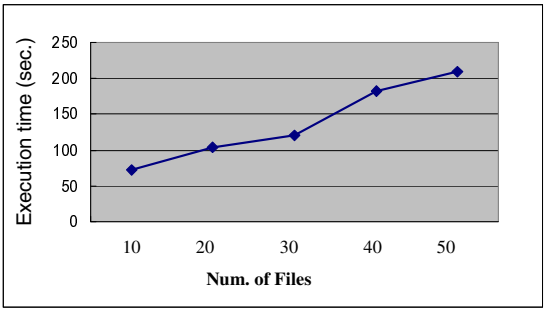


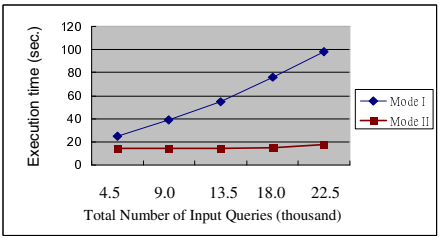
Fig. 5. Execution time of index construction

5.2 Performance of Query Evaluation

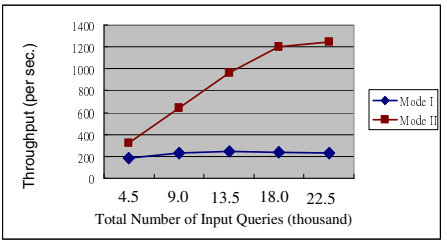
In the query evaluation phase, we feed a large number of queries to the system and examine the performance of query evaluation. The queries used in the experiments are generated by YFilter [18]. The query patterns may be either P-C or A-D relationships. We randomly choose required quantity of queries from the set of distinct queries. Duplicates are allowed in the experiments. We consider two types of query processing modes. Mode I is the streaming mode in which the incoming queries are entered as a stream. Each query is treated as a split and is assigned to a map task. Mode II is the batch mode. A set of queries over the same documents are classified as a group and also treated as a map task. We performed our experiments by using 30 files (1.4GB), a multiple of 15, to maximize the difference between two modes. Fig. 6 (a) shows the results of the execution time on the number of input queries form 4.5 thousand to 22.5 thousand in an increment of 4.5 thousand.

As we mentioned in Subsection 4.2, in Mode I, the system will load corresponding parts of indexes and release them once the end of a query evaluation is reached. The execution time increases steadily with the increase of the number of queries. In Mode II, the system holds the loaded indexes in memory until a group of queries is finished. The indexes can be reused and the cost of I/O is reduced. There is no obvious increase in execution time with the increase in the number of queries for Mode II. The reason is because the time spending on query evaluation is very low comparing to the time spending on I/O. Also because the queries may be duplicated, the loaded indexes in

different tests may be similar. Therefore, the differences in query processing time are not significant. Fig. 6 (b) illustrates the throughput of query evaluation. The throughput is defined as the average number of queries that can be processed in one second. It shows the throughput in Mode I is not improved when the number of queries increases. However, the throughput in Mode II increases rapidly due to the batch processing of the queries.



(a) Execution time of query evaluation



(b) Throughput of query evaluation

Fig. 6. Performance comparisons of query evaluation

6 Conclusions

In this paper, we proposed a system that builds indexes and processes enormous amount of queries for a large number of XML documents using Hadoop framework. The suitability of NCIM, which was developed by our research team, for large number of XML documents is demonstrated in this paper. The experimental results show that the proposed system can deal efficiently with large input XML files. The experimental results also show the throughput of the batch query processing mode is much higher than the streaming mode. In the batch processing mode, the throughput of 1200 queries per second for huge amount of queries using a 15-node cluster signifies the potential applications of NCIM to the fast query processing of enormous Internet documents.

References

1. Liao, I.-E., Hsu, W.-C., Chen, Y.-L.: An Efficient Indexing and Compressing Scheme for XML Query Processing. In: Zavoral, F., Yaghob, J., Pichappan, P., El-Qawasmeh, E. (eds.) NDT 2010. CCIS, vol. 87, pp. 70–84. Springer, Heidelberg (2010)
2. Dutta, H., Kamil, A., Pooleery, M., Sethumadhavan, S., Demme, J.: Distributed Storage of Large Scale Multidimensional Electroencephalogram Data using Hadoop and HBase. In: Grid and Cloud Database Management. Springer, Heidelberg (2011)
3. Thiébaut, D., Li, Y., Jaunzeikare, D., Cheng, A., Recto, E.R., Riggs, G., Zhao, X.T., Stolpestad, T., Nguyen, C.L.T.: Processing Wikipedia Dumps: A Case-Study comparing the XGrid and MapReduce Approaches. In: 1st International Conference on Cloud Computing and Services Science (2011)

4. Goldman, R., Widom, J.: DataGuides: Enabling Query Formulation and Optimization in Semistructured Databases. In: 23rd International Conference on Very Large Data Bases, pp. 436–445 (1997)
5. Al-Khalifa, S., Jagadish, H.V., Koudas, N., Patel, J.M., Srivastava, D., Wu, Y.: Structural Joins: a Primitive for Efficient XML Query Pattern Matching. In: 18th IEEE International Conference on Data Engineering, pp. 141–152. IEEE Press, Washington, DC (2002)
6. Bruno, N., Koudas, N., Srivastava, D.: Holistic Twig Joins: Optimal XML Pattern Matching. In: 2002 ACM SIGMOD International Conference on Management of Data, pp. 310–321. ACM Press, New York (2002)
7. Chen, S., Li, H.G., Tatemura, J., Hsiung, W.P., Agrawal, D., Candan, K.S.: Twig^{2}Stack: Bottom-Up Processing of Generalized Tree-pattern Queries over XML Documents. In: 32nd International Conference on Very Large Data Bases, pp. 283–294 (2006)
8. Qin, L., Yu, X.J., Ding, B.: TwigList: Make Twig Pattern Matching Fast. In: 12th International Conference on Database Systems for Advanced Applications, pp. 850–862 (2007)
9. Pan, Y., Lu, W., Zhang, Y., Chiu, K.: A Static Load-Balancing Scheme for Parallel XML Parsing on Multicore CPUs. In: 7th IEEE International Symposium on Cluster Computing and the Grid, Brazil (2007)
10. Lu, W., Chiu, K., Pan, Y.: A Parallel Approach to XML Parsing. In: 7th International Conference on Grid Computing, pp. 28–29. IEEE Press, Washington, DC (2006)
11. Pan, Y., Zhang, Y., Chiu, K.: Simultaneous Transducers for Data-Parallel XML Parsing. In: 22nd IEEE International Parallel and Distributed Processing Symposium (2008)
12. Pan, Y., Zhang, Y., Chiu, K.: Parsing XML Using Parallel Traversal of Streaming Trees. In: Sadayappan, P., Parashar, M., Badrinath, R., Prasanna, V.K. (eds.) HiPC 2008. LNCS, vol. 5374, pp. 142–156. Springer, Heidelberg (2008)
13. Welcome to Apache™ Hadoop™!, `http://hadoop.apache.org/` (retrieved date: June 27, 2011)
14. Map/Reduce Tutorial, `http://hadoop.apache.org/common/docs/r0.20.2/mapred_tutorial.html` (retrieved date: June 27, 2011)
15. Welcome to Hadoop™ Distributed File System!, `http://hadoop.apache.org/hdfs/` (retrieved date: June 27, 2011)
16. Wikipedia, Apach Hadoop, `http://en.wikipedia.org/wiki/Apache_Hadoop` (retrieved date: June 29, 2011)
17. Zhang, C., De Sterck, H., Aboulnaga, A., Djambazian, H., Sladek, R.: Case Study of Scientific Data Processing on a Cloud Using Hadoop. In: Mewhort, D.J.K., Cann, N.M., Slater, G.W., Naughton, T.J. (eds.) HPCS 2009. LNCS, vol. 5976, pp. 400–415. Springer, Heidelberg (2010)
18. YFilter: Filtering and Transformation for High-Volume XML Message Brokering, `http://yfilter.cs.umass.edu/code_release.html` (retrieved date: June 29, 2011)

Constraint-Based Charging Scheduler Design for Electric Vehicles*

Hye-Jin Kim, Junghoon Lee**, and Gyung-Leen Park

Dept. of Computer Science and Statistics, Jeju National University
690-756, Jeju City, Jeju Do, Republic of Korea
{hjkim82,jhlee,glpark}@jejunu.ac.kr

Abstract. This paper proposes an efficient charging scheduler for electric vehicles and measures its performance, aiming at reducing peak power consumption while satisfying the diverse constraints specified in each charging request. Upon the arrival of a charging request via the underlying vehicle network, the scheduler builds the feasible schedule based on the activation time, the deadline, and the power load profile of each charging task, which is practically nonpreemptive. During the search space expansion of a backtracking algorithm, each step checks the constraint imposed on peak load, completion time, number of chargers, and precedence relation between tasks to prune unnecessary branches. The performance measurement result obtained from the prototype implementation reveals that the proposed scheme reduces the execution time by 80 %, achieves the peak load reduction by 11 %, and improves the schedulability by 5 %, compared with uncoordinated and list scheduling schemes for the given parameter set.

1 Introduction

The smart grid is the next generation power network which combines information technology and the legacy power network to optimize the energy efficiency [1]. It can also make it possible to exchange information on power generation and consumption between those parties, bringing the era of *prosumer*, which means any individual can be both consumer and producer of energy at the same time. While many countries are trying to take initiative in the smart grid, the Korean national government has launched the smart grid testbed in Jeju area, pursuing 5 goals consisting of smart power grid, smart place, smart transportation, smart renewable energy, and smart electricity service [2]. Among these, the smart transportation accelerates the deployment of electric vehicles, which can reduce the CO_2 emissions and noise pollution, consuming much less energy. However, as they need to be charged more frequently than gasoline vehicles due to limited

* This research was supported by the MKE (The Ministry of Knowledge Economy) Korea, under IT/SW Creative research program supervised by the NIPA (National IT Industry Promotion Agency) (NIPA-2011-(C1820-1101-0002)).

** Corresponding author.

J.-S. Pan, S.-M. Chen, N.T. Nguyen (Eds.): ACIIDS 2012, Part III, LNAI 7198, pp. 266–275, 2012.

battery capacity, the smart transportation infrastructure must provide electric charging stations along the road network and at other diverse spots [3].

Even though many researchers and developers are working to improve driving range while decreasing charging time, weight, and cost of batteries, it still takes tens of minutes to charge an electric vehicle [4]. Moreover, drivers want to have their cars charged by a certain time instant, for example, before they depart for their offices, their homes, and the like. As a result, the charging station must schedule multiple requests having different time constraints, charging amount, and power consumption dynamics. In addition, when the station is charging multiple vehicles at the same time, the power consumption may grow too much instantly beyond the permissible bound contracted with its power provider. This situation not just jeopardizes the electricity supply system but also leads to the generation of expensive dispatchable energies. Moreover, the power network may suffer from instant turbulence in the case of unexpected load spikes. Hence, the peak power reduction is important in both economic and environmental aspects as more power plants must be built, if the system-wide peak load exceeds the current power capacity [5].

In this regard, this paper is to design an efficient charging scheduler for the electric vehicles in charging stations. The scheduler mainly finds an appropriate schedule which can reshape the power demand and thus minimize the peak power consumption, meeting the time constraint of each request. Other constraints are also integrated into the backtracking-based scheduler to account for the limitation in the number of chargers as well as the precedence relation between two vehicles. Here, SAE J1772 defines the standard for electric connectors and their charging system architecture, covering physical, electrical, communication protocol, and performance requirement [6]. Our scheduler can work with this standard as a core service for electric vehicles.

This paper is organized as follows: After issuing the problem in Section 1, Section 2 describes the background and related work of this paper. Section 3 defines the service scenario and designs the scheduling scheme. The performance measurement results are discussed in Section 4. Finally, Section 5 summarizes and concludes this paper with a brief introduction of future work.

2 Background and Related Work

Smart transportation is one of the most important areas in the smart grid. Electric vehicles need nation-wide power charging infrastructure, possibly creating a new business model embracing diverse vehicle types, charging stations, and subsidiary services [7]. A charging facility can be installed not just in public stations. We can consider the service area such as universities, offices, public institutes, shopping malls, airport parking lots, and the like. Drivers can charge their vehicles at those places. Noticeably, while a car is being charged, the driver can work at his office or take shopping at a mall. In those places, many vehicles will be concentrated and they must be served according to a well-defined reservation and scheduling strategy [8]. Moreover, as electric vehicles are necessarily

equipped with one or more vehicle network interfaces, they can easily interact with a charging scheduler and other sophisticated services which may reside even in a computing cluster [9].

As for the charging service for electric vehicles, [10] has proposed a reservation-based scheduling scheme for a charging station to decide the service order of multiple requests, aiming at meeting the diverse constraints of as many electric vehicles as possible. Its mechanism makes it possible for a customer to reduce the charging cost and waiting time, while a station can extend the number of clients it can serve. A linear rank function is defined based on estimated arrival time, waiting time bound, and the amount of needed power, for the purpose of reducing the scheduling complexity. Receiving the requests from the clients, the power station decides the service order by the rank function and then replies to the requesters with the waiting time and cost it can guarantee. Each requester can decide whether to charge at that station or try another station. However, $O(1)$ heuristic sacrifices schedule accuracy in the service time and the number of serviceable vehicles. As the charging scheduler can generally afford to tolerate up to tens of seconds, we have time margin to pursue a better solution. That is, it is possible to serve more vehicles, reducing the peak load.

In addition, a power consumption scheduler can minimize the peak load also in individual homes and buildings [11]. To cope with the uncontrollable execution time in finding an optimal schedule, a heuristic is designed for the preemptive task, whose search space complexity is originally estimated to be $O(M^{\frac{M}{2}})$, where M is the number of time slots in the scheduling window. Following the task model consist of actuation time, operation length, deadline, and a consumption profile, the scheduler first investigates all the feasible allocations for nonpreemptive tasks. Next, for each partial allocation, the scheduler calculates the current load of each slot, select slots having the smallest load, and assigns to the preemptive task one by one. For the tasks whose power request does not change significantly over the time slots just like electric vehicle charging, it can achieve reasonable accuracy and optimality in terms of peak load reduction just with an extremely low computation overhead.

3 Scheduling Scheme

3.1 Service Scenario

In our service scenario, a driver tries to make a reservation at a charging station via the vehicular network, while he or she is driving, specifying its requirement details, as shown in Figure 1. Each requirement specifies vehicle type, estimated arrival time, the desired service completion time (deadline), charging amount, and so on. Receiving the request, the scheduler prepares the power load profile of the vehicle type from the well-known vehicle specification. Then, it checks whether the station 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. Entering the station, the vehicle is assigned a charger

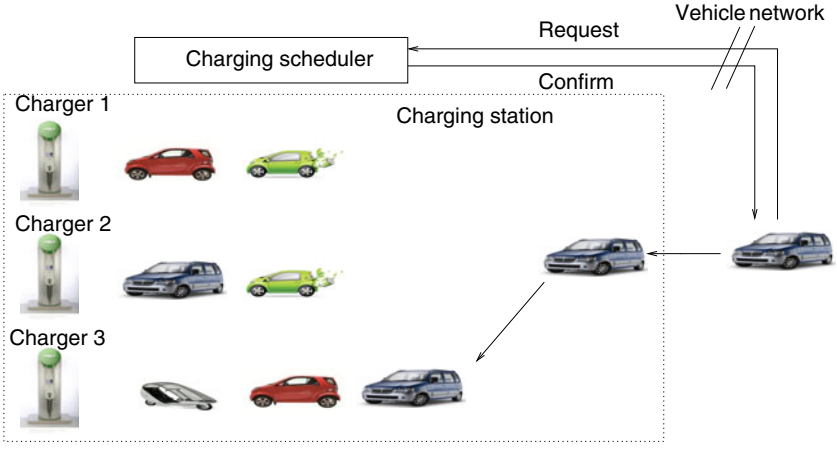


Fig. 1. Service scenario

and waits in the queue according to the schedule. It may take a few minutes for the transition of two vehicles in each charger, during which no vehicle can be served. The scheduler takes into account this effect in its schedule generation as an overhead constant.

3.2 Charging Scheduler

Each charging operation can be modeled as a task. For a task, the power consumption behavior can vary according to the charging stage, remaining amount, vehicle type, and the like. The load power profile is practical for characterizing the power consumption dynamics along the battery charging stage [12]. This profile is important 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. Here, charging tasks starts only at a time slot boundary for the efficiency of schedule generation. The length of a time slot can be tuned according to the system requirement on the schedule granularity and the computing time. In a power schedule, the slot length can be a few minutes, for example, 5 minutes. This length coincides with the time unit generally used in the real-time price signal.

After all, task T_i can be modeled by the tuple of $< A_i, D_i, U_i >$. If there is no control box which can connect or disconnect the power line to each electric vehicle, a charging task is practically nonpreemptive in charging stations. Even though it can be preempted in the single user case as in an individual home, the charging goes to the end once it has started in the charging station. 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 consumption profile entry. A_i is actually the estimated arrival time of the vehicle. Each task can start from its activation time to the latest start time, which can be calculated by subtracting

	0	1	2	3	4	5	6	7	8	9	10	11		18	19	(A_i, L_i, D_i)
T_1	0	0	2	2	3	3	0	0	0	0	0	0	– – – –			(2, 4, 20)
T_2	0	0	0	0	0	4	4	4	3	3	0	0	– – – –			(5, 5, 20)
T_3	0	3	3	3	3	3	3	0	0	0	0	0	– – – –			(1, 6, 20)
T_4	0	0	0	0	0	1	1	2	1	1	1	0	– – – –			(5, 6, 20)
T_5	0	0	0	5	4	5	0	0	0	0	0	0	– – – –			(3, 3, 20)
T_6	0	0	1	1	1	1	1	1	1	1	0	0	– – – –			(2, 8, 20)

Fig. 2. Task model

U_i from D_i. When a start time is selected, the profile entry is just copied to the allocation table one by one, as the task cannot be suspended or resumed during its operation. The choice option is bounded by M, the number of time slots in the scheduling window, hence total search space size, or the maximum number of feasible schedules, can grow up to M^N, where N is the number of tasks.

Figure 2 shows the sample task set consist of 6 tasks. The power requirement is marked from the activation time of each task. Without any schedule, the peak value reaches 17 at Slot 5. This table also depicts how our scheduler works, and Figure 3 describes the detailed scheduling algorithm. For example, if T_2 begins after Slot 6, the peak power consumption will be cut down. The allocation table consists of $M \times N$ fields. The allocation procedure fills the allocation table from the first row, each row being associated with a task. Basically, the procedure creates the search space for all feasible allocations. When the allocation procedure reaches a leaf, *EvalAlloc*() is called to check whether this complete allocation is better than current best in terms of the maximum power consumption. If so, the current best is replaced. If not a leaf, the allocation proceeds after checking the constraint of the current partial schedule. For T_i, for all possible start time between A_i to $D_i - U_i$, the allocation for $(i+1)$-th row is tried one by one. Here, if the consumption profile for the task is (3, 4, 5, 2), the allocation which selects its start time as 2 will be (0, 0, 3, 4, 5, 2, ...) after copying the profile to the table.

CheckConstraint() can speed up the search procedure by pruning the unnecessary search tree expansion. If the maximum power requirement of the partial allocation for the tasks from T_0 to T_i already exceeds the current best, it is no use proceeding to the remaining allocation. In addition, for each time slot, the number of nonzero entries is equal to the number of chargers needed for the schedule. Here, for every column, if the number of nonzero entries exceeds the number of chargers available in the station, the schedule is not valid. Moreover, for the two tasks having a precedence relation, the end time of the precedent task and the start time of the subsequent task are compared. The scheduler can calculate them for the partial or complete allocation table by checking zero-to-nonzero and nonzero-to-zero transitions, respectively. For more speedup, the precedent task is allocated first and the actual activation time of the subsequent task is modified to the completion time of the precedent task. Like this way, a new constraint can be integrated easily.

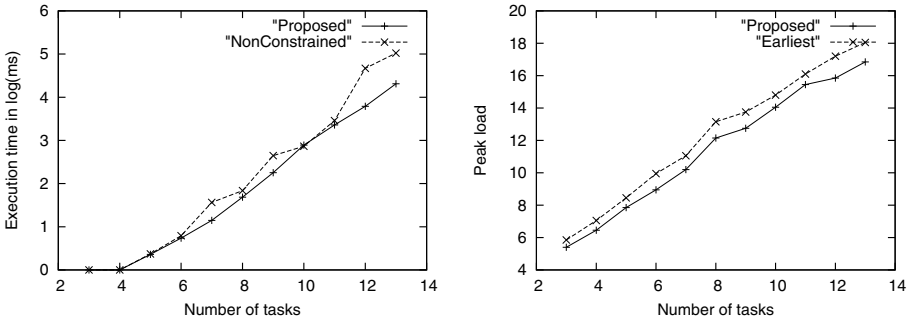
```
procedure CheckConstraint ()
        Check the current peak load of the partial schedule and current best
        Check the number of chargers needed for the allocation
        Check the precedence relation
end procedure

procedure AllocTab (i)
input : {T_i|(D_i, A_i, U_i)}
        if i equals to N
            EvalAllloc ()
        end if
        for each start time from A_i to D_i − U_i
            copy the profile
            if (! CheckConstraint ()) AllocTab (i+1)
end procedure
```

Fig. 3. Vehicle charging schedule

4 Performance Measurement

This section implements the prototype of the proposed allocation method using Visual C++ 6.0. Our implementation runs on the platform equipped with Intel Core2 Duo CPU, 3.0 GB memory, and Windows Vista operating system. The experiment sets the schedule length, namely, M, to 20 time units. If one time unit is equal to 5 min, the total schedule length will be 100 min, but the time scale is a tunable parameter. For a task, the activation time and the operation length randomly distribute between 0 and M. However, we make the finish time, namely, the sum of start time and the operation length, not exceed M. All tasks have the common deadline, namely, M, for simplicity, but this restriction can be removed. In addition, the power demand for each time slot has the value of 1 through 5. The power scale, for example, kw, is not explicitly specified in this experiment, as it is a relative-value term. The execution time is measured using Microsoft Windows $GetTickCount$ system call which has the 1 ms time granularity.

Figure 4 measures the effect of the number of tasks to the scheduling time and peak load reduction. For each number of tasks, 20 sets are generated and their results are averaged. As the time complexity of the scheduling algorithm is estimated to be $O(M^N)$, the computation time is plotted in the log scale. Figure 4(a) reveals the efficiency of constrained processing, where unnecessary branches in the search space are pruned. The performance gap between the constrained search and the nonconstrained exhaustive search increases along with the increase of the number of tasks, reaching 80 % when the number of tasks is 13. Figure 4(b) plots peak load reduction achieved by the proposed scheme, comparing with the *Earliest* scheduling scheme. This scheme initiates the task as soon as the task is ready. In our experiment, the total amount of power demand is proportional to the number of charging tasks. The performance



(a) Execution time (b) Peak power reduction

Fig. 4. Effect of the number of charging tasks

graph indicates that our scheme can efficiently reshape the power consumption pattern, reducing the peak value by 11 %. The peak load reduction increases along with the number of tasks.

In addition to the absolute execution time comparison, Figure 5 plots the number of visited leaves in the search tree in the same parameter set with Figure 4, as it can show the platform-independent performance criteria. A leave corresponds to a complete schedule that has survived branch pruning. Here, the branch pruned in the higher level of the tree more reduces the number of visited leaves. As can be seen in the figure, two curves increases largely proportional to the number of tasks. However, the proposed scheme can reduce the number of visited leaves to 3.6 %, when the number of tasks is 12.

Figure 6 measures the effect of the number of chargers in a station. This experiment fixes the number of tasks to 10, while changing the number of chargers from 3 to 10. As shown in Figure 6(a), the nonconstrained search shows the constant execution time, while the constrained search is affected by the number of chargers. When the number of chargers is equal to the number of tasks,

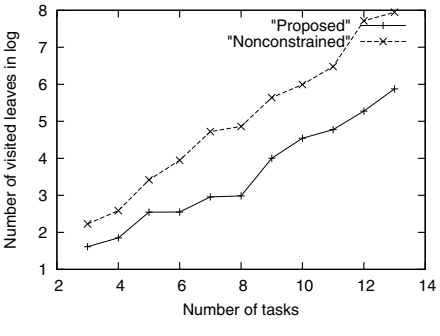
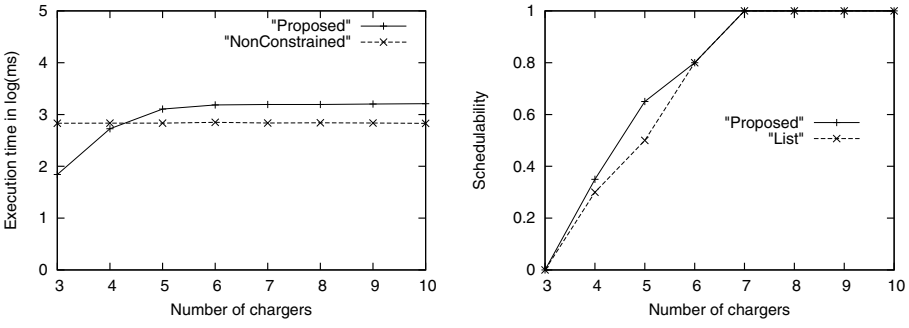


Fig. 5. Analysis of the number of visited leaves



(a) Execution time (b) Accuracy

Fig. 6. Effect of the number of chargers

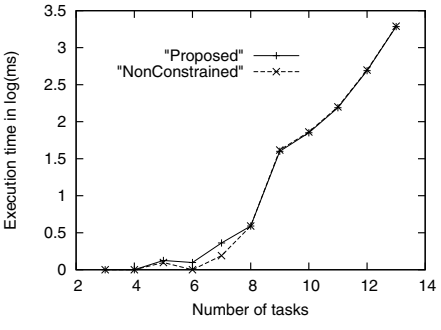


Fig. 7. Effect of precedence constraint

namely, 10, constraint processing is meaningless, just increasing the unnecessary overhead for checking the validity of partial schedules. Hence, the weaker the constraint, the higher the overhead. Figure 6(b) plots the schedulability of the proposed scheme, comparing with *List* scheduling. The basic idea of list scheduling is to make an ordered list of tasks by a specific priority, and then repeatedly select from the list the task with the highest priority and assign to a resource until a valid schedule is obtained. We have generated 20 sets and checked if two scheduling schemes can find an appropriate schedule. The ratio of schedulable sets to the total sets is defined to be schedulability. When the number of chargers is equal to or larger than 7, both schemes can find the schedule for all generated sets. When number of chargers is from 4 to 6, the proposed scheme shows better schedulability by 5 %.

Finally, Figure 7 plots the effect of precedence constraint processing. For each task set having the number of tasks from 3 to 13, Task 1 and Task 2 have precedence constraint. Namely, Task 1 must be completed before Task 2 begins. The first two tasks are selected as they are handled first in the search space expansion, removing more branches at the higher level of the search tree. Basically,

constraint processing can eliminate the search space that violates this constraint. However, for the partial and complete schedules which meet the precedence constraint, it also performs constraint checking. Hence, the execution time for both cases is almost same except the interval from 5 to 8.

5 Concluding Remarks

This paper has proposed an efficient charging scheme for electric vehicles in charging stations and measures its performance, mainly aiming at reducing peak power consumption. Receiving the charging request specifying its own constraint via the underlying vehicle network, the scheduler generates a schedule and sends confirmation back to the requester. While expanding the search space based on the classic backtracking algorithm, it checks the constraint on peak time, completion time, number of chargers, and precedence relation to prune unnecessary branches. The performance measurement result obtained from the prototype implementation reveals that our scheme can reduce the execution time by 80 %, achieves the peak load reduction by 11 %, and improves the schedulability by 5 %, compared with existing earliest and list scheduling schemes. The efficiency in the charging station system can accelerate the deployment of electric vehicles in smart transportation.

As future work, we are first planning to design a charging station selection algorithm for the convenient driving of electric vehicles, and the charging schedule can possibly integrate the renewable energy option [13]. Here, the vehicle battery can be used as power storage. Besides such charging station applications, it is also promising to analyze the variety of sensor data for the electric vehicles, as it can create valuable information for car safety and management [14].

References

1. Gellings, C.W.: The Smart Grid: Enabling Energy Efficiency and Demand Response. CRC Press (2009)
2. Korean Smart Grid Institute, http://www.smartgrid.or.kr/eng.html
3. Guille, C., Gross, G.: A Conceptual Framework for the Vehicle-to-grid (V2G) Implementation. Energy Policy 37, 4379–4390 (2009)
4. Markel, T., Simpson, A.: Plug-in Hybrid Electric Vehicle Energy Storage System Design. In: Advanced Automotive Battery Conference (2006)
5. Spees, K., Lave, L.: Demand Response and Electricity Market Efficiency. The Electricity Journal, 69–85 (2007)
6. Sanzhong, B., Yu, D., Lukic, S.: Optimum Design of an EV/PHEV Charging Station with DC Bus and Storage System. In: IEEE Energy Conversion Congress and Exposition, pp. 1178–1184 (2010)
7. Kaplan, S.M., Sissine, F.: Smart Grid: Modernizing Electric Power Transmission and Distribution; Energy Independence, Storage and Security. TheCapitol.Net (2009)
8. Schweppe, H., Zimmermann, A., Grill, D.: Flexible In-vehicle Stream Processing with Distributed Automotive Control Units for Engineering and Diagnosis. In: IEEE 3rd International Symposium on Industrial Embedded Systems, pp. 74–81 (2008)

9. Ipakchi, A., Albuyeh, F.: Grid of the Future. IEEE Power & Energy Magazine, 52–62 (2009)
10. Kim, H.-J., Lee, J., Park, G.-L., Kang, M.-J., Kang, M.: An Efficient Scheduling Scheme on Charging Stations for Smart Transportation. In: Kim, T.-H., Stoica, A., Chang, R.-S. (eds.) SUComS 2010. CCIS, vol. 78, pp. 274–278. Springer, Heidelberg (2010)
11. 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)
12. Tremblay, O., Dessaint, L.: Experimental Validation of a Battery Dynamic Model for EV Applications. World Electric Vehicle Journal 3 (2009)
13. Caramanis, M., Foster, J.M.: Management of Electric Vehicle Charging to Mitigate Renewable Generation Intermittency and Distribution Network Congestion. In: 48th IEEE Conference on Decision and Control, pp. 4717–4722 (2009)
14. Midlam-Mohler, S., Ewing, S., Marano, V., Guezennec, Y., Rizzoni, G.: PHEV Fleet Data Collection and Analysis. In: IEEE Vehicle Power and Propulsion Conference, pp. 1205–1210 (2009)

Modular Arithmetic and Fast Algorithm Designed for Modern Computer Security Applications

Chia-Long Wu*

Professor and Director of Aviation Communication Electronics Department
Chinese Air Force Institute of Technology
chialongwu@gmail.com

Abstract. Modular arithmetic plays very crucial role for public key cryptosystems, such as the public key cryptosystem, the key distribution scheme, and the key exchange scheme. Modular exponentiation is a common operation used by several public-key cryptosystems, such as the RSA encryption scheme and the Diffie-Hellman key exchange scheme. In this paper, we have proposed a new method to fast evaluate modular exponentiation, which combines the complement recoding method and canonical recoding technique.

Keywords: Canonical recoding, modular arithmetic, complexity analyses, fast algorithm design, public-key cryptosystem.

1 Introduction

Modular exponentiation is the fundamental operation in implementing circuits for cryptosystem, as the process of encrypting and decrypting a message requires modular exponentiation which can be decomposed into multiplications. In this paper, a proposed multiplication method utilizes the complement recoding method and canonical recoding technique. By performing complement representation and canonical recoding technique, the number of partial products can be further reduced. Exponentiation is a basic yet important operation for public key cryptography. In this paper, an efficient modular exponentiation method is proposed by adopting the binary method, common-multiplicand multiplication, complement method and signed-digit recoding method. Hamming weight plays an important part for complexity efficiency. On average, by performing minimal Hamming recoding method and signed-digit recoding method, the number of multiplications for our proposed algorithm can be reduced effectively, where k is the bit-length of the exponent E. We can therefore efficiently speed up the overall performance of the modular exponentiation [1].

To compute modular exponentiation $C \equiv M^E \bmod N$, (where C, M, E, and N are ciphertext, plaintext, public key, and modulus respectively) is very time-consuming

* Corresponding author.

J.-S. Pan, S.-M. Chen, N.T. Nguyen (Eds.): ACIIDS 2012, Part III, LNAI 7198, pp. 276–285, 2012.

because the bit-length of E can be up to 2048 bits. Designing efficient algorithms that can speed up software and hardware implementation of modular exponentiation are often considered as practical significance for practical cryptographic applications such as the RSA public-key cryptosystem [1] and the ElGamal cryptosystem [2].

Speeding up modular exponentiation $C \equiv M^E \bmod N$, where $E = \sum_{i=1}^{k} e_i \times 2^i$ and $e_i \in \{0,1\}$, is very crucial for public key cryptosystems (PKC). There are several well-known algorithms for speeding up the exponentiations and the multiplications such as binary exponentiation method (sometimes called the square-and-multiply method) [3], signed-digit recoding method [4-11], exponent-folding-in-half method [12-13], Montgomery reduction method [14-15], common-multiplicand multiplication (CMM) method [16-19], and multi-exponentiation method [20-21], and so on.

The Hamming weight (the number of 1's in the binary representation) plays an important role for the computational efficiency. A novel method for speeding up modular exponentiation by using binary exponentiation method, complement recoding method, and signed-digit recoding method is proposed in this paper. We can efficiently speed up the overall performance of modular exponentiation.

The rest of this paper is organized as follows. Some related methods are introduced in Section 2. In Section 3, the proposed algorithm for fast modular exponentiation is described. Then, the computational complexity of the proposed algorithm is analyzed in Section 4. Finally, we conclude this work and future works in Section 5 [28-30].

2 Mathematical Preliminaries

2.1 The Binary Exponentiation Method

Fast computations of the exponentiation can be classified into two approaches: the faster multiplication designs, and the development of novel exponentiation algorithms. The multiplication involves two basic operations, the generation of partial products and their accumulation. The binary exponentiation method [3] also called square-and-multiply method is a generally acceptable method for exponentiation. It can convert the modular exponentiation of $C \equiv M^E \bmod N$ [23] into a sequence of modular multiplications. Let the exponent E have the binary representation $E = \sum_{i=1}^{k} e_i \times 2^i$, where $e_i \in \{0, 1\}$ and k is the bit- length of the exponent E.

It can be divided into two kinds of methods. One is the right to left binary exponentiation method; the other is the left to right binary exponentiation method. The right to left binary exponentiation method scans the exponent E from the least significant bit (LSB) toward the most significant bit (MSB). It performs one multiplication operation and one square operation when the exponent bit e^i is 1 and performs one square operation when the exponent bit e^i is 0. It will be shown as Algorithm 1[28-30].

Algorithm 1. Right to Left binary exponentiation algorithm

Input: Message: M;
Exponent: $E = (e^k e^{k-1} \ldots e^2 e^1)2$;
Output: Ciphertext: $C = M^E$;
begin
$C = 1$;
$S = M$;
for $i = 1$ to k do /*scan from right to left */
{ if ($e^i = 1$) then C = C×S; /*multiply*/
S = S×S;} /*square*/
endfor
end.

The left to right binary exponentiation method scans the exponent E from the most significant bit (MSB) toward the least significant bit (LSB). It performs one multiplication operation when the exponent bit e^i is 1 and performs one square operation when the exponent bit e^i is 0. It will be shown as Algorithm 2 [25-27].

Algorithm 2. Left to Right binary exponentiation algorithm

Input: Message: M;
Exponent: $E = (e^k e^{k-1} \ldots e^2 e^1)_2$;
Output: Ciphertext: $C = M^E$;
begin
C = 1;
S = M;
for $i = k$ to 1 do /*scan from left to right */
{C = C×C; /*square*/
if ($e^i = 1$) then C = C×S;} /*multiply*/
endfor
end.

The computational complexity of both algorithms expresses as follows. On an average, we assume the occurrence probabilities for both bit "1" and bit "0" are the same i.e. {S×S} and {S×S, C×S} with the same probability. Then, the expectation value for bits "1" and "0" is the same "$\frac{k}{2}$", where k is the bit-length of the exponent E.

2.2 The Bit-Complement Recoding Method

To compute the modular exponentiation of $C \equiv M^E \bmod N$, we express the exponent E as a binary representation $e_k e_{k-1} \ldots e_2 e_1$. Performing complements is advantageous in the speed up of exponential computations [24-26]. The equation and example will be shown as Equation 1.

$$E = \sum_{i=1}^{k} e_i \times 2^i = (e^k e^{k-1} \ldots e^2 e^1)_2 = (10\ldots0)_{(k+1)\text{bits}} - \overline{E} - 1, \quad (1)$$

where $\overline{E} = \overline{e_k}\,\overline{e_{k-1}} \ldots \overline{e_1}$ and $\overline{e_i} = 0$ if $e_i = 1$; $\overline{e_i} = 1$ if $e_i = 0$, for $i = 1, 2, \ldots, k$.

2.3 The Signed-Digit Recoding Method

In a signed-digit number with radix 2, three symbols $\{\bar{1}, 0, 1\}$ are allowed for the digit set, in which 1 and $\bar{1}$ in bit position i represented $+2^i$ and -2^i respectively [3]. It shows that the average Hamming weight of a k-bit canonically recorded binary number approaches $\frac{k}{3}$ as $k \to \infty$ [4-5, 22]. We should note that a number using the digit $\{\bar{1}, 0, 1\}$ is not uniquely represented in binary signed-digit notation [6]. The equation and example will be shown as Algorithm 3 [25-29].

Algorithm 3. Signed-Digit Recoding Method

Input: E = $(e^k e^{k-1} \ldots e^2 e^1)_2$;
Output: E_{SD}
begin
c1 = 0; rn+2=0; rn+1=0;
for i = 1 to k do
$c_{i+1} = \left\lfloor \frac{c_i + e_i + e_{i+1}}{2} \right\rfloor$;
$e_i = c_i + e_i - 2c_{i+1}$
endfor
return E_{SD}
end.

Since the addition of k bits can generate an integer with magnitude of $\log(k)$ bit addition, the cost needs only $\frac{\log(k)}{k}$ k-bit additions. Here "$A<<<8$" stands for the integer which is obtained by the left-shift eight bits from the multiplicand A. Since the Hamming weight of multiplier B is larger than $\frac{k}{2}$, the Hamming weight of $\overline{B}$ is $(\frac{k}{2} * \frac{1}{2}) = \frac{1}{4}k$ in average, i.e., $A * \overline{B}$ needs $\frac{k}{4}$ k-bit additions. Therefore, we need two $2k$-bit subtractions. Assume that both addition and subtraction have the same computational complexity [25-28].

2.4 The Common-Multiplicand Multiplication Method

In 1993, Yen and Laih proposed the common-multiplicand multiplication (CMM) method to improve the performance of the right-to-left binary exponentiation algorithm for evaluating modular exponentiation "$M^E \bmod N$". Here we concentrate on the computations of $\{A \times B_i | i = 1, 2, \ldots, t;\ t \geq 2\}$. The following variables are required in the CMM method (for $i = 1, 2, \ldots, t$) [25-27],

$$B_{com} = B_1 \textbf{ AND } B_2 \ldots \textbf{AND } B_t, \tag{2}$$

$$B_{i,c} = B_i \textbf{ XOR } B_{com}, \tag{3}$$

where "**AND**" and "**XOR**" are bitwise logical operators.

Hence B_i can be depicted as:

$$B_i = B_{i,c} + B_{com} \quad \text{for } i = 1, 2, \ldots, t. \tag{4}$$

Therefore, the common-multiplicand multiplications $A \times B_i$ (i = 1, 2,…, t) can be computed with the assistance of $A \times B_{com}$ as:

$$A \times B_i = A \times B_{i,c} + A \times B_{com} \quad \text{for } i = 1, 2, \ldots, t. \tag{5}$$

In 1961, Avizienis proposed a signed-digit (SD) representation, also called redundant number representations for parallel and high-speed arithmetic. A signed-digit vector representation of an integer a in radix r is a sequence of digits $a = (a_{k+1}, a_k \ldots, a_2, a_1)_{\text{SD}}$ with $a_i \in \{0, \pm 1, \cdots, \pm r-1\}$ for $k \geq i \geq 0$, i.e., $a = \sum_{i=1}^{k+1} a_i \times r^i$. In a binary signed-digit number (BSD) system, three symbols {$\bar{1}$, 0, 1} are allowed for the digit set, the symbol $\bar{1}$ is used to denote the value -1[28-30].

The basic idea of CMM method is to extract the common parts of multiplicands, and save the number of binary additions for the computation of common parts. Let A and B_is (i = 1, 2) be m-bit integers, the Hamming weights of B_i, B_{com} and $B_{i,c}$ are $m/2$, $m/2^t$ and ($m/2 - m/2^t$), respectively. By using the CMM method, the computations of $\{A \times B_1, A \times B_2\}$ can be represented as $\{A \times B_{1,c} + A \times B_{com}, A \times B_{2,c} + A \times B_{com}\}$.

The total number of binary additions for the common-multiplicand multiplications evaluation is $m/2^t + t \times (m/2 - m/2^t)$. Without the CMM method, the multiplications $\{A \times B_1, A \times B_2\}$ are computed one after another independently using total $t \times (m/2)$ binary addition. Thus, the performance improvement of the common-multiplicand multiplication method shown above can be denoted as [28-30]:

$$\frac{\frac{mt}{2}}{\frac{m}{2^t} + t \times (\frac{m}{2} - \frac{m}{2^t})} = \frac{t}{t + (1-t) \times 2^{1-t}}. \tag{6}$$

The auxiliary carry C_0 is set to 0 and subsequently the binary number A is scanned two bits at a time. The canonically recoded digit Bi and the next value of the auxiliary binary variable C_{i+1} for i = 0, 1, 2, … , n are generated as shown in Table 1.

Table 1. Canonical recoding table

A_{i+1}	A_i	C_i	B_i	C_{i+1}
0	0	0	0	0
0	0	1	1	0
0	1	0	1	0
0	1	1	0	1
1	0	0	0	0
1	0	1	$\bar{1}$	1
1	1	0	$\bar{1}$	1
1	1	1	0	1

3 The Proposed Signed-Digit Recoding Algorithm

In Section 2, we describe the binary exponentiation method, complement recoding method, and signed-digit recoding method respectively. We combine these methods as Algorithm 5 [28-30] to accelerate the exponentiation. Algorithm 4 is the signed-digit recoding method and is depicted as follows).

Algorithm 4. Signed-Digit Recoding Algorithm

Input: Message: M;
Exponent: E = $(e^k e^{k-1} \ldots e^2 e^1)_2$;
Output: Ciphertext: $C = M^E$;
begin
C = 1;
S = M;
for i = 1 to k do /*scan from right to left */
{if $(e_i = 1)$ then $C \equiv (S \times C) \bmod N$; /*multiply*/
if $(e_i = \bar{1})$ then $C \equiv (S^{-1} \times C) \bmod N$;
$S \equiv (S \times S) \bmod N$. /*square*/
end.

Algorithm 5. The Proposed Signed-Digit Recoding Algorithm

Input: Message: M;
Exponent: $E = (e^k e^{k-1} \ldots e^2 e^1)_2$;
Modulus: N;
Output: Cipher-text: $C \equiv M^E \bmod N$
begin
Count the Hamming weight of E, denote as Ham(E).
if Ham(E)> $\frac{k}{2}$
Perform the complement recoding and the signed-digit recoding procedures.
C = 1;
S = M^{-1};
for i=1 to k do
{if $(e_i = 1)$ then $C \equiv (S \times C) \bmod N$;
if $(e_i = \bar{1})$ then $C \equiv (S^{-1} \times C) \bmod N$;
$S \equiv (S \times S) \bmod N$; }
else
Perform the signed-digit of E is E_{SD}.
Call Signed-Digit Binary algorithm (M, E, N): C;
Output C;
end.

4 The Complexity Analyses of the Proposed Algorithm

In this section, we will describe the computational complexity of the proposed algorithm. The computational complexity of the proposed method is $\frac{1}{3}k+2*\frac{\log(k)}{k}+5\approx 0.333k$ k-bit additions that are faster than $\frac{3}{4}k\approx 0.75k$ in Yen-Laih method, $\frac{23}{32}k\approx 0.719k$ in Wu-Chang method, $\frac{7}{12}k\approx 0.583k$ in Yen's method and $\frac{1}{2}k+2*\frac{\log(k)}{k}+5\approx 0.5k$ in Chang- Kuo-Lin method. Here are the complete complexity analyses.

We assume there are k bits in exponent E. There are two cases [28-30]:

Case 1: Ham(E)>$\frac{k}{2}$ and Case 2: Ham(E)$\leq\frac{k}{2}$.

The computational complexity of $C\equiv M^E \bmod N$ = (the computational complexity of Step 1)+ ($\frac{1}{2}\times$ the computational complexity of Step 2) + ($\frac{1}{2}\times$ the computational complexity of Step 3).

The second and the third items "$\frac{1}{2}$" in the above equation mean the probabilities of Ham(E)>$\frac{k}{2}$ and Ham(E)$\leq\frac{k}{2}$.

Assume that the multiplicand A and multiplier B are k-bit unsigned binary numbers. The computational complexity of $P=A*B$ is defined as " (the computational complexity of Step 1) + ($\frac{1}{2}\times$the computational complexity of Step 2) + ($\frac{1}{2}\times$the computational complexity of Step 3)". The second and the third items "$\frac{1}{2}$" in the above equation mean the probabilities of Ham(B) > $\frac{k}{2}$ and Ham(B) $\leq\frac{k}{2}$.

Now we describe the computational complexity of Step 1, Step 2(Case 1) and Step 3 (Case 2) respectively. First, we define E_{SD} a binary signed-digit representation for E and $\overline{E_{SD}}$ a binary signed-digit representation for $\overline{E}$ respectively.

Then, the computational complexity is counted on the number of k-bit multiplication [28-30].

Step 1: scan E from LSB to MSB

We scan E from the least significant bit (LSB) toward the most significant bit (MSB) to sum them up and check if Ham(E)>$\frac{k}{2}$. The computational complexity of this step is much less than that of multiplication [28-30].

Step 2: Ham(E)>$\frac{k}{2}$

We consider 1's complement of E as $\overline{E}$, i.e. Ham $(\overline{E})<\frac{k}{2}$. We can replace $C\equiv M^E \bmod N$ by $C\equiv M^{(10...0)_{(k+1)\text{bits}}-\overline{E}-1} \bmod N\equiv M^{(10...0)_{(k+1)\text{bits}}-\overline{E_{SD}}-1} \bmod N$. On an

average, the Hamming weight of $\overline{E_{\mathrm{SD}}}$ is $\frac{k}{2}\times\frac{1}{3}=\frac{k}{6}$, where "$\frac{1}{3}$" is non-zero digit probability for $\overline{E_{\mathrm{SD}}}$ by using signed-digit recoding method [28-30].

$$\begin{aligned} C &\equiv M^{E} \bmod N \equiv M^{(1111100001)_2} \bmod N \\ &\equiv M^{(10..0)_{(k+1)\text{bits}}-\bar{E}-1} \bmod N \equiv M^{(10000000000)_{11\text{bits}}-0000011110-1} \bmod N \\ &\equiv M^{(10..0)_{(k+1)\text{bits}}-\overline{E_{\mathrm{SD}}}-1} \bmod N \equiv M^{(10000000000)_{11\text{bits}}-000010000\bar{1}0-1} \bmod N \\ &\equiv (M^{(10000000000)_{11\text{bits}}} \times (M^{-1})^{000010000\bar{1}0} \times M^{-1}) \bmod N \end{aligned} \tag{7}$$

where k is the bit-length of the exponent and $E=(1111100001)_2$.

5 Conclusions

In this paper, we have proposed a fast method to efficiently evaluate modular multiplication, which combines the complement recoding method and canonical recoding technique [29-30]. The computational complexity of the proposed method is faster than Yen-Laih method, Wu-Chang method [12], Yen's method [17] and in Chang- Kuo-Lin method [24-28]. We can efficiently speed up the overall performance of multiplication operation by using the proposed algorithm.

As the modular squaring operation in finite field can be done by a simple shift operation when a normal basis is used, and the modular multiplications and modular squaring operations in our proposed signed-digit recoding scheme can be executed in parallel, by using our proposed generalized r-radix signed-digit folding algorithm, hardware design and parallel technique, we can effectively decrease the computational complexity [27-30].

References

1. Rivest, R.L., Shamir, A., Adleman, L.: A method for obtaining digital signatures and public key cryptosystems. Communications of the ACM 21(2), 120–126 (1978)
2. ElGamal, T.: A public key cryptosystem and a signature scheme based on discrete logarithms. IEEE Transactions on Information Theory 31(4), 469–472 (1985)
3. Knuth, D.E.: The Art of Computer Programming, 3rd edn. Seminumerical Algorithms, vol. 2. Addison-Wesley, MA (1997)
4. Yang, W.C., Guan, D.J., Laih, C.S.: Algorithm of asynchronous binary signed-digit recording on fast multi-exponentiation. Applied Mathematics and Computation 167(1), 108–117 (2005)
5. Koc, C.K., Johnson, S.: Multiplication of signed-digit numbers. Electronics Letters 30(11), 840–841 (1994)
6. Avizienis: Signed-digit number representations for fast parallel arithmetic. IRE Transactions on Electronic Computers 10, 389–400 (1961)
7. Arno, S., Wheeler, F.S.: Signed digit representations of minimal Hamming weight. IEEE Transactions on Computers 42(8), 1007–1010 (1993)

8. Syuto, M., Satake, E., Tanno, K., Ishizuka, O.: A high-speed binary to residue converter using a signed-digit number representation. IEICE Transaction on Information and Systems E85-D(5), 903–905 (2002)
9. Heuberger, C., Prodinger, H.: Carry propagation in signed digit representations. European Journal of Combinatorics 24(3), 293–320 (2003)
10. Joye, M., Yen, S.M.: Optimal left-to-right binary signed-digit recoding. IEEE Transactions on Computers 49(7), 740–748 (2000)
11. Koren: Computer Arithmetic Algorithms, 2nd edn. A. K. Peters, MA (2002)
12. Lou, D.C., Chang, C.C.: Fast exponentiation method obtained by folding the exponent in half. Electronics Letters 32(11), 984–985 (1996)
13. Lou, D.C., Wu, C.L., Chen, C.Y.: Fast exponentiation by folding the signed-digit exponent in half. International Journal of Computer Mathematics 80(10), 1251–1259 (2003)
14. Montgomery, P.L.: Modular multiplication without trial division. Mathematics of Computation 44(170), 519–521 (1985)
15. Tenca, F., Koc, C.K.: A scalable architecture for modular multiplication based on Montgomery's algorithm. IEEE Transactions on Computers 52(9), 1215–1221 (2003)
16. Yen, S.M., Laih, C.S.: Common-multiplicand-multiplication and its applications to public key cryptography. Electronics Letters 29(17), 1583–1584 (1993)
17. Yen, S.M.: Improved common-multiplicand-multiplication and fast exponentiation by exponent decomposition. IEICE Transaction on Fundamentals E80-A(6), 1160–1163 (1997)
18. Wu, T.C., Chang, Y.S.: Improved generalization common-multiplicand-multiplications algorithm of Yen and Laih. Electronics Letters 31(20), 1738–1739 (1995)
19. Ha, C., Moon, S.J.: A common-multiplicand method to the Montgomery algorithm for speeding up exponentiation. Information Processing Letters 66(2), 105–107 (1998)
20. Dimitrov, V.S., Jullien, G.A., Miller, W.C.: Complexity and fast algorithms for multi-exponentiations. IEEE Transactions on Computers 49(2), 141–147 (2000)
21. Chang, C.C., Lou, D.C.: Parallel computation of multi-exponentiation for cryptosystems. International Journal of Computer Mathematics 63(1-2), 9–26 (1997)
22. Wu, C.-L., Lou, D.-C., Lai, J.-C., Chang, T.-J.: Fast modular multi-exponentiation using modified complex arithmetic. Applied Mathematics and Computation 186(2), 1065–1074 (2007)
23. Stallings, W.: Cryptography and Network Security Principles and Practice, 3rd edn. Prentice-Hall, NY (2002)
24. Chang, C.C., Kuo, Y.T., Lin, C.H.: Fast algorithms for common multiplicand multiplication and exponentiation by performing complements. In: Proceeding of 17th International Conference on Advanced Information Networking and Applications, pp. 807–811 (March 2003)
25. Wu, C.L.: Fast modular multiplication based on complement representation and canonical recoding. In: The 7th Conference of Crisis Management (CMST 2009), Tainan, Taiwan, pp. 1–8 (November 27, 2009)
26. Wu, C.L.: Modular exponentiation arithmetic and number theory for modern cryptographic security applications. In: 8th Conference of Crisis Management (CMST 2010), CCM1010002IFS, Kaohsiung, Taiwan, pp. 169–176 (2010)
27. Wu, C.L.: High performance of modular arithmetic and theoretical complexity analyses. In: Proceedings of the 7th Pacific Symposium on Flow Visualization and Image Processing (PSFVIP-7), pp. 18–35 (November 2009)

28. Wu, C.L., Lou, D.C., Chang, T.-J., Chen, C.-Y.: Fast modular exponentiation algorithm theoretical design and numerical analysis for modern cryptographic applications. In: 17th National Defense Science Technology Symposium (ND17), Taoyuan, Taiwan, November 27-28, vol. 5-1–5-7 (2008)
29. Wu, C.-L.: Complexity analyses and design for cryptographic modular algorithm. In: 2011 Symposium on Communication Information Technology on Management and Application, Paper No. 0505, Kaohsiung, A2: Communication Theory, pp. 1–6 (2011)
30. Wu, C.-L.: Fast Montgomery binary algorithm for information security. In: 2011 International Symposium on NCWIA, Paper No. 111, Kaohsiung, D6: Information Systems and Innovative Computing, pp. 1–5 (2011)

An Efficient Information System Generator

Ling-Hua Chang[1] and Sanjiv Behl[2]

[1] Kun Shan University of Technology, Department of Information Management
No. 949, Da Wan Rd., Tainan City, Taiwan, R.O.C.
[2] Thomas Edison State College,
101 W. State St, Trenton, NJ 08608-1176
changlh@mail.ksu.edu.tw, sanbehl@yahoo.com

Abstract. We developed a new customized software tool for automatically generating a complete Java program based on the values or parameters inputted by the user. We call it an efficient Information System Generator or ISG for short. It is efficient in terms of the processor usage and the development time. We illustrate how it can be used by building a system for keeping track of student's scores that can be used by any faculty member who teaches multiple courses at a university or a college. It can also be used for generating e-commerce web sites.

Keywords: Java GUI Generator, prototype, Information System Generator, E-commerce Generator.

1 Introduction

We know that building a software system is a very time-consuming process and therefore we designed a customized software tool to help people generate any information (software) system instead of writing it themselves. We call this an Efficient Information System Generator or ISG for short. For example, we can use ISG to generate a conference system generator, which can be used to regenerate any conference system for any company or institute after they understand the context of the whole conference. ISG can also be used to generate e-commerce web sites, for instance the following website can be generated using ISG - http://www.36086789.com/category-2-b0.html?=1013yyyy .

The system designers need to provide the graphical user interface and input data to the ISG. The GUI interface includes how the input is obtained and the output is displayed, as well as the relationship between the two. Once this information is inputted, the ISG tool will convert it to the necessary Java code and database.

2 Related Work

Microsoft Excel [1] is a commercial spreadsheet application written and distributed by Microsoft for Windows and Mac OS X. It has a "*smart recalculation*" feature where when the cell data changes, only the data associated with it will be updated, and the user

J.-S. Pan, S.-M. Chen, N.T. Nguyen (Eds.): ACIIDS 2012, Part III, LNAI 7198, pp. 286–297, 2012.

can immediately see the changes resulting from the change in the cell data. It also has powerful graphics capabilities, so the users can see when the graphic data changes. We hope that in the future ISG would be as simple to use as Excel is today.

There are many Java generators currently on the market like JFlex[2], BOUML[3], Javadoc [4] and JAG[5]. JFlex is a lexical analyzer generator, BOUML generates the code based on the definition made at the artifact, class, operation, relation, attribute and extra member levels. Javadoc is the Java API Documentation and generates HTML pages of API documentation from Java source files. JAG (Java Application Generator) is an excellent tool and uses open source Apache Ant, a Java library and a command-line tool. The system design provides the database, object, window frame and files information, which can be generated to an information system that is built on the J2EE platform. However we think that it is too difficult to use for most small and medium enterprises.

UJECTOR [6] is a tool for creating executable code from UML Models. It uses UML class, sequence and activity diagrams for automated code generation. It generates structural code from the class diagram, and then adds behavioral aspects from the sequence and activity diagrams. In the rule based production systems for automatically generating Java code (or RPSAGJ) [7], the user writes the requirements in simple English and the designed system is able to extract associated information after compound analysis, which is then used to draw various UML diagrams as activity, sequence, class and uses cases diagrams. The designed system has a robust ability to create code automatically without external environment.

ISG is also a development tool for generating an information system which incorporates the advantages of Excel and JAG and also introduces a prototyping output design. We illustrate how our tool can be used to design and prototype the input and the output layout for the system users and programmers. UJECTOR and RPSAGJ are consistent with the UML models, however, the current version of ISG creates colorful, attractive and user-friendly screens which can be used by an information system or for generating any e-commerce web system. At present we are using ISG to develop a business information system for East Land International Company. The screens provided by UJECTOR and RPSAGJ are not very convenient or user-friendly for entering data as they are in ISG. We are looking into incorporating their features in the future versions of ISG.

3 ISG System Architecture

Since the *ISG i*nformation system is a Java GUI tool, the user must provide system analysis and system design information, including the panel design and its functionalities, the relationship between the panels and where the data storage takes place. ISG offers users interface screens which can be used for generating an information system and has seven transformation functions - for building a database, for building files, for linking to the next window, for building data processing window, for displaying data, for previewing a designed window and for printing data. We now discuss each transformation function individually.

3.1 Building a Database

ISG provides a function for building a database from a large collection of data that allows users to search for and extract the needed information.

3.2 Building Files

ISG can load the Java Swing components or the data objects onto the memory and then translate it into Java programs for an information system. For example, consider a system for keeping track of the student's scores in a course. Each teacher may teach 4 or 5 courses in a semester and needs to keep track of the student's scores for all the assignments and tests in a course. Therefore this system needs to provide the following functions – to enable the entry of courses, the percentage rate of scores, a description of the test scores, students' names, homework or quiz scores, midterm scores and final scores; and to show the score list and the students' final grades. The data is stored using

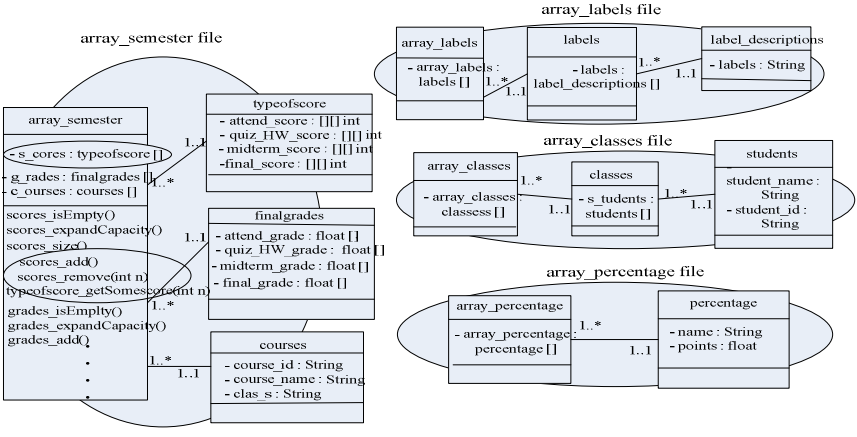


Fig. 1. Class diagram for the file structure of the Student Scores system

object streams rather than regular streams. Because Java has *persistence* in object-oriented circles [8], that means we can let the object layout on disk to be exactly like the object layout in memory. So we save an object (the first created object) to disk and objects that are created subsequently, whose memory addresses are stored in each array, are managed and stored to the disk automatically. ISG file system introduces this mechanism. For using this mechanism, we designed the file structure of the system as shown in Fig 1. There are 4 object stream files in the system, each of which is shown as an oval in the figure and labeled at the top of it – array_semester file, array_labels file, array_classes file and array_percentage file. There are 3 arrays in the array_semester file, which are shown below the top label in the array_semester rectangle, for storing scores, grades and courses. Array s_cores is used to store students' scores for each class in this semester (shown in a separate rectangle), array

g_rades is used to store the student's final grades for each class, and array c_ourses is used to store the course number and name for each course. An element of array s_cores is an object of class typeofscore that contains four different arrays viz. attend_score, quiz_HW_score, midterm_score and final_score array. They are used to store the student's scores for the various components that make up the final score or grade. Array_percentage file stores information on the type of test like the midterm, final, or homework-quiz, and the percentage weight of each in the final grade. Array_labels file is used to describe the details of each test type such as chapter 1 homework assignment or quiz on March 22 and those scores are counted in the quiz and homework scores. Array_class file stores the student names and student ids.

Consider the array_semester file for illustrating how the object streams are stored in a file. ISG creates an object. For example, the array_semester is an object type. Related to it are objects typeofscore, finalgrades and courses which are stored in s_cores, g_rades and c_ourses arrays respectively. Method add() is used to add an object to its associated object array. For example, scores_add() adds a typeofscore object to the s_cores array. Method remove() removes an object from its associated object array. For example, scores_remove(int n) removes a typeofscore object from the array s_cores. Method *typeofscore_getSomescore*(int n) is used to get a typeofscore object from the array s_cores. Thus the file structure of Student scores system, with the arrays and the methods provided makes it convenient to retrieve the data elements in any array.

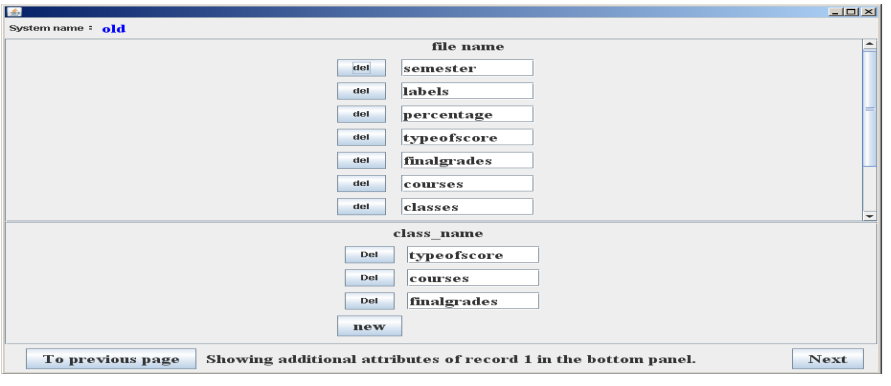


Fig. 2. Building each file of the system

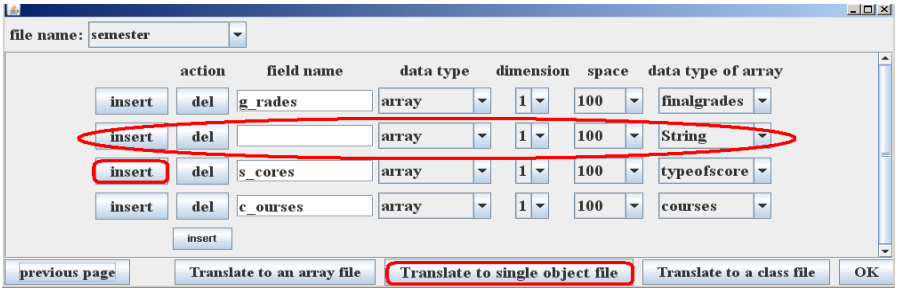


Fig. 3. Building each field of every class

We now discuss how ISG builds the file system of the student scores system. Fig 2 shows a shot of the window that is used to create all the files and their objects for the system, for the classes and files shown in Fig 1 viz. array_semester, typeofscore, finalgrades, courses, array_labels, labels, label_descriptions, array_classes, classes, students, array_percentage and percentage. Fig 3 is used to set the attributes of each class. In the screen shot shown, the attributes of class *semester* are being set, which from Fig 1 are g_rades, s_cores and c_ourses. The figure shows that each of these attributes is a one dimensional array of the same size, and the data type of the arrays is finalgrades, typeofscore and courses respectively. When the user presses the *Translate to a single object file* button, array_semester.java program file will be created. Thus the file structure shown in Fig 1 and the screen shots shown in Fig 2 and 3 can be used to generate the Java classes that have the add, remove, etc methods which can be used to manage an array.

3.3 Linking to the Next Window and Building a Data Processing Window

There are two types of windows provided in the ISG - a link to the next window (Fig 4) and building a data processing window (Fig 5). Fig 4 shows the homepage of the student's scores system and has 5 buttons and one text field. The name of the directory where the data is saved is entered into the textfield by the user. The figure shows the directory as "Fall of 2011", so all the data will be saved in a directory by that name. Clicking on any of the buttons will take you to the corresponding window. Fig 5 shows the records being entered by entering the course id and the course name in the respective fields. In addition to that, *object serialization* mechanism [8] is used to fill objects with data, which saves CPU execution time. These ideas were used to design the following four different kinds of windows to manage data:

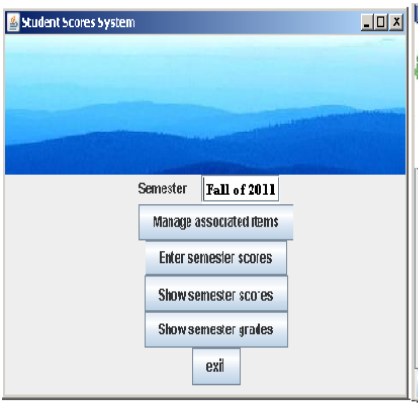


Fig. 4. A link to next window

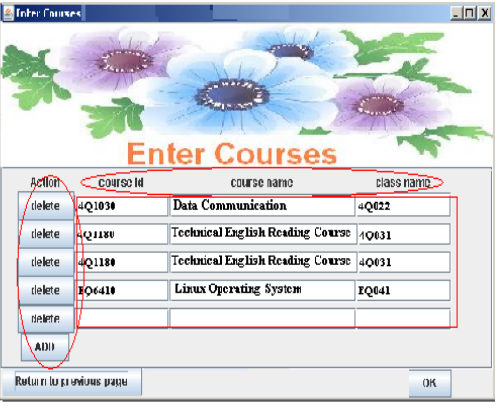


Fig. 5. A basic data processing window

1. *Basic data processing window* (Fig 5) allows for the data of the same type to be added, modified and deleted. Can use the add button to add a new record, and the delete button to delete the corresponding record. Pressing the OK button saves the data to an object stream file.
2. *Basic data processing window plus more data attributes* (Fig 6) has all the features of the *basic data processing window* and shows more data attributes in a panel at the square bottom. When the number of data attributes is more than can be shown in the basic data processing window, then the additional attributes of a record can be shown in the bottom panel. In the figure shown, the message at the bottom of the window says "showing additional attributes of record 2 in the bottom panel", which means that the focus is on an attribute of record 2 and the additional attributes of that record are being shown in the bottom panel.
3. *Special data processing window* (Fig 3) is obtained by adding insert buttons next to the delete buttons in the *basic data processing window*. Pressing an insert button displays an empty line of fields on top of the line whose insert button was pressed, and can be used to enter a new record.
4. *Special data processing window plus more data attributes* - when you add a panel at the bottom of the special data processing window for entering more attributes, then you get this window.

Figures 6, 7 and 8 show the attributes that need to be set in order for the ISG to translate it to a Java GUI program. For example, consider the window shown in Fig 5, which lets a teacher enter the courses in a semester. Before ISG translates it, the user needs to enter the window size and location on the screen, the fonts for displaying text and images, etc. This information is needed by the Java layout manager for creating the window. Fig 6 displays the window where the users can enter each window's name, frame title, frame size, window location, window type and where the data was read from and where the data will be written to. The oval mark in the figure shows the name of the window to be *course*, frame title to be *Enter courses*, frame size to be 500 by 500, location to be (0,0) and type of window to be *basic data processing window*.

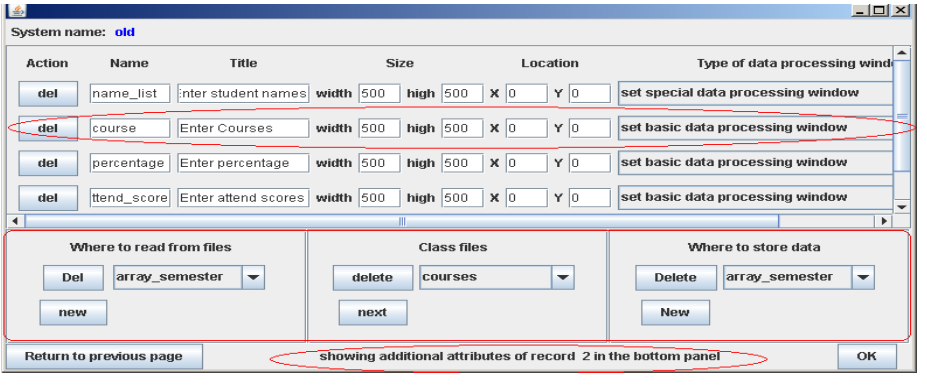


Fig. 6. All the data processing windows of students' score system will be set

At the bottom of the Fig 6, there are three columns; *where to read from files*, *class files* and *where to store data*. The *where to read from files* specifies the files the ISG needs to read from. After users update *Enter Courses* data, the data will be written back to the files specified by *where to store data*. *Class files* tells ISG where each data attribute such as course id, course name and class name, is located. In Fig 8, the rectangle on the right hand side shows the class diagram of the stored file array_semester. Because ISG builds each file structure as an object stream file and array will manage objects of the same class, therefore in the class diagram, it shows

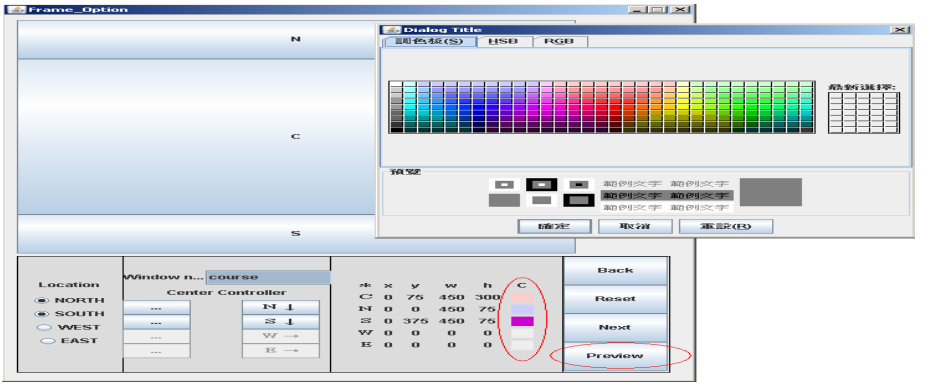


Fig. 7. Selecting a panel to set its' attributes

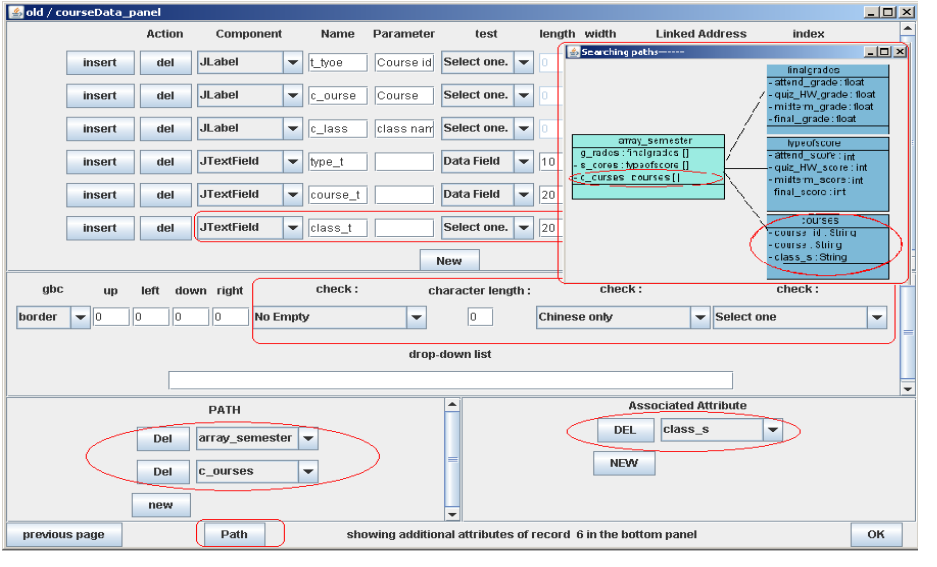


Fig. 8. Set Swing components of *data processing window* course

that every object stored in the file will be stored in its associated array. Therefore the objects created by class finalgrades are stored in array g_rades, the objects created by class typeofscore are stored in array s_cores and the objects created by class coures are stored in array c_ourses because the array_semester file stores student scores, final grades and courses. In the *Enter Courses* window (Fig 5), the course id, course name and class name entered by the users are stored in an object of class *courses.* This class and its' attributes course_id, course and class_s are shown in the rectangle on the right hand side of Fig 8. Therefore in Fig 6, class *courses* should be selected in *class files* column, in the *where to read from files* column semester should be selected and in the *where to store data* column semester also should be selected. Fig 7 is used to select panels located on the window - north, south, east or west panel. The center panel is always there but the other four panels are selectable. ISG offers a color space for users to set a background color for each chosen panel (see area marked with an oval with a 'c' on the top of Fig 7). Fig 8 is used to set the Swing components of the *Enter courses* window (Fig 5). The labels at the top are implemented by using JLabels components, and the textfields for entering the attributes of a record are implemented by using JTextfield components. See the oval at the top in Fig 5, there are 3 labels in it viz. course id, course name and class name.

For the labels, we set the names and parameters of the JLabel components (t_type and course id for the first one, and so on). For the text fields, we set the names and sizes of the JTextfield components (type_t and 10 for the first one, and so on). We can also specify constraints on the data that can be entered into the text fields. For example, we can specify that the text field for entering the class name has to be not empty, and only chinese values can be entered in it. This can be accomplished by setting the parameters shown inside the rectangle shown near the middle of the figure. *Associated Attribute* (bottom right, Fig 8) is used to specify where in the file each data will be stored. Since ISG file structure is an object stream file, each entry data will be stored in an object of some class. In Fig 5, when users enter course id, course name and class name, they actually type in the following text fields viz. type_t, course_t and class_t separately. The data they enter is encapsulated into an object of class *courses* with attributes course_id, course and class_s. Therefore the value of associated attribute course_id is set to what is entered in textfield type_t, course is set to value entered in textfield course_t, and class_s is set to value entered in textfield class_t. The Associated Attribute rectangle only shows class_t. ISG knows how to get and set these attributes since they are entered by the user, but doesn't know where to start searching from until it finds the value of the needed attribute class_s. That information can be specified in the Path rectangle (bottom left of the figure). The Fig 8 specifies the path of the course_id and course or class_s.

We mentioned earlier that ISG builds object stream files using Java persistence which means that the object layout on a disk is exactly like the object layout in memory. Therefore we save an object to disk and then the memory addresses of the objects that are created subsequently are stored in each array. For example, ISG creates an object of class array_labels and then stores many memory addresses of objects of class labels in array array_labels. Later, there are many objects of class label_descriptions are created and then store their memory addresses in array labels.

However if the file system is complicated, it would not be easy for the users to remember it. Therefore ISG provides a class diagram of stored files to help users to set

the PATH and Associated Attribute. For our student's courses system example, the class diagram of array_semester file will be displayed (see the right top rectangle in Fig 8) when the Path button is pressed. From the class diagram, we know course id, course name and class name are stored in an object of class courses and its associated attribute is course_id, course and class_s. Therefore these objects will be stored in an array c_ourses and that is one of arrays stored in an object of class array_semester. PATH (in the left bottom of Fig 8) is used for ISG to translate how to retrieve attribute course_id, attribute course and attribute class_s easily. The class diagram shows only class array_semester and class course needed in this window. Therefore we know the object array_semester is stored in file and in this object contains 3 arrays. There is only array courses what we need and which store objects of class courses with values of course_id, course and class_s. Therefore we need to get object array_semester first and then each object with course_id, course and class_s will be stored in array c_ourses. Therefore, users will set path to array_semester first and then c_ourses next (see in Fig 8). ISG got these values of path and then it is every easy for ISG to translate it into a data processing window course.java and also includes where and how to input or output these data. The advantage of this idea is the file structure introduced is very useful for ISG to retrieve every attribute of an object with users helping to offer these parameters.

3.4 Displaying and Printing the Data

There are three sets of data to be displayed for the Student scores system viz. quiz or homework_scores, midterm scores and final scores. Query data is displayed in a table and can be sorted. Users can press the print button to print the table.

3.5 Previewing a Designed Window

Pressing the next button in figure 7 will translate all the Swing components of the already set *data processing window course* (see Fig 8) to Java GUI programs and then pressing the preview button will compile and execute these programs and then show the designed *data processing window* on the screen (see Fig 5). The users can go back and forth through the windows by pressing the next and previous page buttons. When the whole system is done, it can be shown to the users to give them an idea of what the system might be like or how it will work. If they have any new ideas or suggestions or if this prototype is rejected, the feedback can be used to modify it. This cycle can be repeated until an acceptable or desired system is created. Can press the reset button to go back to the original values or the original state and can start over again.

4 Experimental Results and Analysis

4.1 Improving CPU Efficiency

The CPU time is one of the most important resources and determines how fast a program executes. Therefore we used a Linked queue [9] data structure to manage the GUI, which lets us easily create a window by adding, deleting, or modifying the Swing

components to the window. Linked queue (see figures 9, 10) is a data structure that uses object reference variables to create links between objects and is large enough to hold the numeric address of an object. It is a dynamic data structure because its size grows and shrinks as needed to accommodate the number of elements stored.

For example, consider the four classes of students we discussed earlier. We use an array to represent the four school classes and then another array *classes* to represent students of each class. Because an array is a convenient data structure for storing objects and each array element implementations is efficient, we only allocate enough space per element for the object reference variable. If managed properly, by using an appropriate initial capacity and then expanding it as needed, this additional space is not a problem. Therefore we use an array to store a collection.

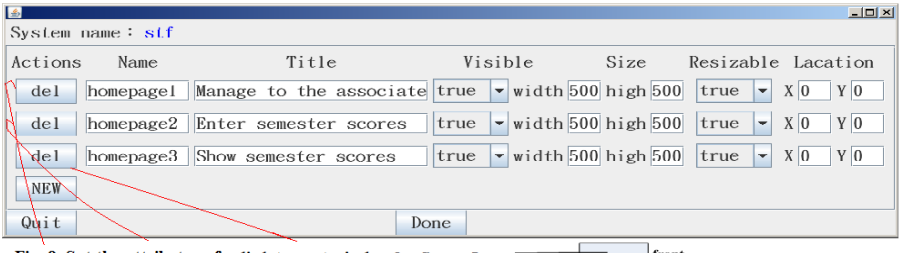


Fig. 9. Set the attributes of a *link to next window* for figure 5

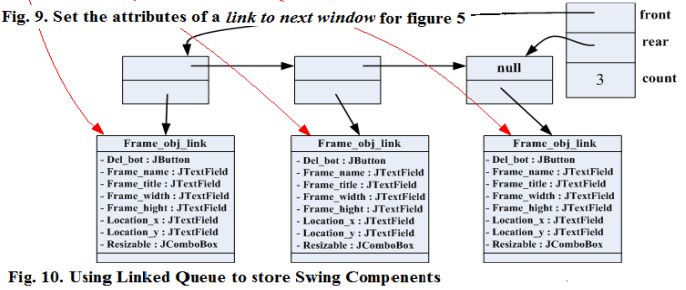


Fig. 10. Using Linked Queue to store Swing Compenents

4.2 ISG has a Quick Development Process Time

ISG helps to develop a prototype quickly which allows the objectives to be tested and developed even further. ISG uses the concept of a spiral model [10] in which feedback from the earlier prototype is used to further refine it. Thus ISG is a user-friendly and efficient program generating software tool which can be used to generate Java programs.

ISG can be used to generate a prototype quickly which can then be shown to the system designers and users. When they see this prototype and start using the system, they might find some functions that does not meet their needs, or certain functionality that is missing. That information can be used by the developers to revise the system to meet users' requirements. Since changes can be done quickly by using our ISG, it can help users understand what they need and what the information system should be like, thus resulting in an end product which would be much better than what it would be otherwise. Since ISG can save time in writing, debugging and testing the program, it would also lower the cost of producing software written in Java.

5 Conclusions and Future Work

We showed how our efficient information system generator was used to develop a system for keeping track of student's scores by generating a total of 56 Java GUI programs. ISG is also very efficient since, for instance, it translated 300 lines of code in the system we illustrated in just 32 milliseconds.

We are currently developing an I-Conference generator that can generate JSP code for a conference system. A conference is a meeting of people who "confer" about a topic. When you need to develop a conference system for your department, school, institute or company for example, you can use ISG to help you implement it after you do the analysis and design. I-Conference generator is a GUI interface for translating the input provided by the user to JSP or HTML programs. The user provides the name of each web page and their corresponding documents or information resources on the Web. After the Web pages have been created, the user needs to specify what the data (in the text fields of the Web pages) is about and where to store it (such as in a database table or a file). The generator will then convert it to the appropriate JSP code, which can be stored on a server and the conference can be installed on a Web site. The GUI interfaces can be completely generated by ISG regardless of how many windows are needed.

Our ISG is not 100% ready at the moment for generating all of the GUI windows for an I-Conference. We still need to add more functions to the *data processing windows*. Since the four types of data processing windows we discussed use a fixed-length record format, we need to change that as records are rarely of the same length. Therefore we need to add different length records format for building the *data processing windows*.

We are in collaboration with East Land International Company Limited for developing a business information system for them. It is an international business company that exports glass containers and health food. We will be developing an information system for them that would involve computing the monthly shipping amount, generating reports on their monthly earnings, profit, orders, etc.

References

1. Excel (2010), `http://office.microsoft.com/en-us/excel/`
2. JFlex- The Fast Scanner Generator for Java, `http://jflex.de/`
3. BOUML- a free UML 2 tool box, `http://bouml.free.fr/doc/javagenerator.html`
4. javadoc - Java API Documentation Generator, `http://download.oracle.com/javase/1.4.2/docs/tooldocs/windows/javadoc.html`
5. JAG- Java Application Generator, `http://jag.sourceforge.net/`
6. Usman, M., Nadeem, A., Kim, T.-H.: UJECTOR: A tool for Executable Code Generation from UML Models. IEEE Advanced Software Engineering & Its Applications, 165–170 (2008)

7. Bajwa, I.S., Siddique, M.I., Choudhary, M.A.: Rule based Production Systems for Automatic Code Generation in Java. IEEE Digital Information Management, 300–305 (2006)
8. Horstmann, C.S., Cornell, G.: Core Java Volume I–Fundamentals, 8th edn. Sun Microsystems Press, Prentice Hall (2008)
9. Lewis, J., Chase, J.: Java Software Structures designing and using data structures. Pearson Education Inc. (2005)
10. Whitten, J.L., Bentley, L.D., Dittman, K.C.: System analysis design methods. McGraw-Hill (2004)

A Temporal Data Model for Intelligent Synchronized Multimedia Integration

Anthony Y. Chang

Department of Information Technology, Overseas Chinese University,
Taichung, Taiwan, R.O.C.
achang@ocu.edu.tw

Abstract. This paper aims to develop a generic temporal data model and to design the relevant mining for multimedia applications. We develop a data model as efficient computation model to unify time-varying events and objects. Several computation algorithms and operation tables which include a set of complete temporal logics are proposed. The combined temporal data model is generalized by composing point and interval algebra with qualitative and quantitative functions. The temporal data model is extended to spatial projection representations. Also we apply the result of computation models to multimedia data for synchronized integration applications. This data model framework can handle semantics of the solutions for knowledge querying and can compose the temporal semantics among multimedia objects over the web.

Keywords: Temporal Data Model, Intelligent Systems, SMIL, TTML, Temporal Database.

1 Introduction

A multimedia presentation is playing the various objects composing the scenario on some appropriate output devices after a specification generated. The W3C standard Synchronized Multimedia Integration Language (SMIL) is wildly used in recent internet. It provides to describe multimedia presentation to be distributed over the web[8][9], and allows authors to specify how components are related temporally to media objects or events in an interactive multimedia presentation. SMIL is supported by the QuickTime, Real, and Windows media architectures.

The simplest way to place video advertising around an on-demand clip is to top-and-tail the content with preroll and postroll using a SMIL file to play the clips serially. Windows media services use active server pages to generate dynamic playlists. SMIL can be used to program the playlist. If you are using the Apple QuickTime streaming server, the administration terminal has a user interface for creating and editing playlists. This can be used for MPEG-4 or MP3 playlists.[4]

The W3C standard Timed Text Markup Language (TTML) is a content type that represents timed text media for the purpose of interchange among authoring systems. It is intended to be used for the purpose of transcoding or exchanging timed text

J.-S. Pan, S.-M. Chen, N.T. Nguyen (Eds.): ACIIDS 2012, Part III, LNAI 7198, pp. 298–307, 2012.

information among legacy distribution content formats [10]. It refers to the presentation of text media in synchrony with other media, such as audio and video. Timed text for MPEG-4 movies and mobile media is specified in MPEG-4 Part 17 Timed Text, and its MIME type is specified by RFC 3839.

This extends temporal interval relation by means of a complete analysis for temporal computation. We define a temporal algebra system for unifying and scheduling SMIL multimedia presentations and TTML documents. The mechanism is efficiently to analyze temporal relationships and to generate synchronization scenarios. The authoring tools help uses designing a presentation more easily, also the system automatically encode a designed presentation as a consistent SMIL and TTML documents.

2 Integrated Multimedia Systems

Multimedia application usually contains a number of multimedia resources to be presented sequentially or concurrently. These resources need to be arranged as layout. We design a visualized user interface by using temporal algebra [3] and updated composing algorithms for controlling multimedia synchronization. The spatio-temporal derived engine constructs partial order relations about temporal and spatial relations as the core of the intelligence distributed presentation system compatible with SMIL and TTML modules. Give an overview of the system:

- Multimedia Resource Browser: before a user is about to design a presentation, multimedia resources are indicated from web allocation.
- Temporal Specification Editor: the presentation system allows authors to specify how components are related temporally to media objects or events in an interactive multimedia presentation.
- Spatial Specification Editor: the spatial relations are extended from temporal relations, to specify the spatial information by 2-tuple relations.
- Synchronization and Layout Engine: computes schedules and layouts between the resources, and translates algebra representation into SMIL and TTML documents or inverse.
- Presentation Database: a multimedia document database for grouping application or content exchange in distributed environment.

3 Temporal Representation

Representation and Reasoning about temporal information play an important role in current computer science. The work discussed in [1] analyzes the relations among temporal intervals. However, the work [1] only states temporal interval relations and point relation was not discussed.

We denote by *before*, *equal* and *after*(denoted as {<}, {=} and {>}) for representing relations between two points. Qualitative variables take these values only. The notation R1 and R2 denotes the point relations over three points A, B, and C which A R1 B and B R2 C. The meaning of some qualitative operators is defined as follows. Some values are uncertain and denoted by $\top$.

Table 1. Point addition

R1 \ R2	<	=	>
<	<	<	T
=	<	=	>
>	T	>	>

In order to express more precise relations without losing qualitative information, the temporal relations extended with qualitative mechanisms for handling quantitative information. To give a concrete form to the topic of temporal representation, consider the following variable and equations with quantitative and qualitative information.

Definition 1: Formal Endpoint Relations.

A formal endpoint relation $Q_E = (E_R, V_E)$ is a quantitative-qualitative valuable.

where E_R, is an endpoint relation based on the point space {<, =, >}, and V_E is a quantity which expresses a quantitative value associated with E_R between two end-points. ■

Let (R1, v1) and (R2, V2) denotes two formal endpoint relations, R1 and R2 are two qualitative temporal variables, v1 and v2 are two quantity associated with quality. The meaning of some quantitative-qualitative calculus operators and equality are defined as follows.

Table 2. Qulitative-quantitative addition

(R1,v1) \ (R2,v2)	$<_{v2}$	=	$>_{v2}$
$<_{v1}$	$<_{v1+v2}$	$<_{v1}$	$<_{v1-v2}$, if (v1>v2) =, if (v1=v2) $>_{v2-v1}$, if (v1<v2)
=	$<_{v2}$	=	$>_{v2}$
$>_{v1}$	$>_{v1-v2}$, if (v1>v2) =, if(v1=v2) $<_{v2-v1}$, if(v1<v2)	$>_{v1}$	$>_{v1+v2}$

Table 3. Qulitative-quantitative subtraction

(R1,v1) \ (R2,v2)	$<_{v2}$	=	$>_{v2}$
$<_{v1}$	$<_{v1-v2}$, if (v1>v2) =, if (v1=v2) $>_{v2-v1}$, if (v1<v2)	$<_{v2}$	$<_{v1+v2}$
=	$<_{v2}$	=	$>_{v2}$
$>_{v1}$	$>_{v1+v2}$	$>_{v1}$	$>_{v1-v2}$, if (v1>v2) =, if(v1=v2) $<_{v2-v1}$, if(v1<v2)

Table 4. Qulitative-quantitative minus

(R1, v1)	[-](R1, v1)
$<_{v1}$	$>_{v1}$
=	=
$>_{v1}$	$<_{v1}$

In addition, a quantitative-qualitative equation correctly expresses both qualitative equation and quantitative equations by formal endpoint variables and operators. We give a set of equations, for an example:

- Given $[d_A]$, $[d_B]$ and [AsBs]
  $[AsBe] = [AsBs] + [d_B]$
  $[AeBs] = -[d_A] + [AsBs]$
  $[AeBe] = -[d_A] + [AsBs] + [d_B]$
- Given $[d_A]$, $[d_B]$ and [AsBe]
  $[AsBs] = [AsBe] - [d_B]$
  $[AeBs] = -[d_A] + [AsBe] - [d_B]$
  $[AeBe] = -[d_A] + [AsBe]$

where $[d_A]$, $[d_B]$, [Ab], [Bb], [Ae], and [Be] are expressing duration of A, duration of B, begin of A, begin of B, end of A, and end of B respectively. ■

4 Parsing Multimedia Documents

In order to identify multimedia documents and to play a multimedia presentation we first parse the document into phrases based on tags. SMIL 3.0 defines major functional grouping of attributes. The timing and synchronization module sets are the core of SMIL specification. We focus on timing, synchronization, media control as an authoring tool for helping user analyze temporal synchronization. After parsing a variable-length SMIL code, each temporal identifier is translate to consistent algebra notation and reasoning about complete temporal relations. A time-line scheduling is also provided for assistant of a presentation designing. The semantics of the SMIL and TTML documents could be extracted from temporal relations of the beginning and ending point. It is also represented as a high level interval form.

With group manipulation, group objects of a presentation could be reused and be played independently. User could search presentation document, media attribute, and semantics from presentation database.

Since the SMIL and TTML documents is base on timeline model with a dynamic time graph, the inconsistency often occurs in both qualitative semantics and quantitative values. We propose a methodology based on point algebra to deal with qualitative and quantitative inconsistency. A temporal scenario is a set of temporally events. There are three basic time containers in a SMIL presentation document. The <seq>

container plays a sequence of children in which elements play one after the other. The <par> container plays a group of children in which multiple elements can share a common timebase and playback at the same time. The <excl> container plays one child at a time, but only one of the children can be active at the time.

Modeling a temporal scenario often requires synchronizing the distributed multimedia objects. In this section, an integrated temporal computational model is constructed to deal with the following inconsistencies.

- **Qualitative inconsistency**: the conflicts occur in the semantics or logics of temporal relations. Such as a scenario plays dependently with other objects which can not be satisfied with temporal relations.
- **Quantitative inconsistency**: the conflicts occur in the scheduling on syncbase-value, event-value, or offset-value.
- **Resources limitation**: resources are needed to accomplish scenario actions, they are limited and obliged to coordinate actions to avoid each other. Resources have to be shared with eliminating pointless actions, avoiding any possible conflicts.
- **Mutual Constraints:** the conflicts occur in the interdependencies between actions of scenario.

5 Intelligent Scheduling for Multimedia Synchronization

The Synchronization and Layout Engine could translate the temporal algebra into SMIL and TTML or inverse. The following examples show how timing could be encoded with both.

Table 5. SMIL and temporal algebra representations

SMIL Specifications	Algebra Notations
<seq> <text id="A" src="a.txt" begin="5s"/> </seq>	[Ab]=5s [Ae] <10 [Bb]
<par> </par>	[Cb]=5s [Ce]=10s [Db]>$_2$[Cb] [Db]=[Cb]+2s=7s

A TTML document instance referenced from a SMIL presentation is expected to follow the same timing rules as apply to other SMIL media objects [10].

Quantitative-qualitative physics is concerned with the dynamic behavior of the physical word. Some temporal models just reason about qualitative or quantitative constraints and some models process qualitative and quantitative constraints separately. Using the point/interval composition and quantitative-qualitative equations, the quantitative-qualitative composition functions could be constructed.

Figure 1 gives an example to represent an interval relation with endpoint constraints qualitatively. Given two intervals A and B, if end of A is before end of B, i.e.

Ae✧Be = {<}, we can derive As✧Bs = {T}, As✧Be = {<}, Ae✧s = {T} by point transitivity, and conclude A could "before", "overlaps", "meets", "during", or "starts" B (denoted as {<, o, m, d, s}).

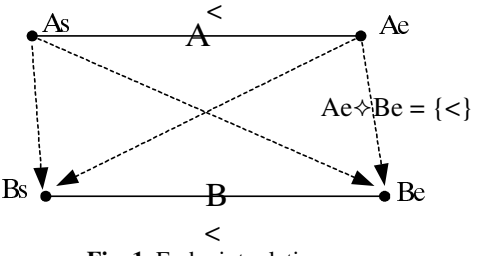


Fig. 1. Endpoint relations

Definition 2: An *integrated temporal algebra* or is defined by the 6-tuple

$\boldsymbol{T} = (d_1, d_2, Rss, Rse, Res, Ree)$

where

d_1 , $d_2 \in Q_E$ called the duration. It contains a fixed quality {<} and a quantity to label the quantitative value between starting point and ending point.

Rss , Rse, Res, $Ree \in Q_E$ contains the endpoint relation Ab✧Bb between N_1 and N_2, and a quantity associating with endpoint relation Ab✧Bb. ■

Definition 3: Given a nonempty set $\boldsymbol{T} = (d_1, d_2, Rss, Rse, Res, Ree)$, ⊗ is a binary operation on $\boldsymbol{T}$, ⊗: $\boldsymbol{T} \times \boldsymbol{T} \rightarrow \boldsymbol{T}$ is the quantitative-qualitative composition function. ■

Theorem 1: Quantitative-Qualitative Composition Functions

Let (dA, dB, AbBb, AbBe, AeBb, AeBe) ⊗ (dB, dC, BbCb, BbCe, BeCb, BeCe) =(dA, dC, AsCs, AsCe, AeCs, AeCe) then

[AbCb] = [AbBb] + [BbCb] = [AbBe] + [BeCb]
[AbCe] = [AbBb] + [BbCe] = [AbBe] + [BeCe]
[AeCb] = [AeBbs] + [BbCb] = [AeBe] + [BeCb]
[AeCe] = [AeBb] + [BbCe] = [AeBe] + [BeCe] ■

Example 1: Considering three temporal interval of actions, A, B, and C, following requirements are just be known:

- The duration of A is 20-units length.
- The duration of B is 10-units length.
- The duration of C is 16-units length.
- Beginning of A is before beginning of B for 30 units.
- End of B is after beginning of C for 13 units.

In integrated temporal algebra, the information could be denoted as:

$[d_A] = <_{20}$
$[d_B] = <_{10}$
$[d_C] = <_{16}$
$[AsBs] = <_{30}$
$[BeCs] = >_{13}$

The complete temporal knowledge is derived after following derivation.

Deriving formal endpoint relations:

$[AsBe] = [AsBs] + [d_B] = <_{30} + <_{10} = <_{40}$
$[AeBs] = -[d_A] + [AsBs] = >_{20} + <_{30} = <_{10}$
$[AeBe] = -[d_A] + [AsBs] + [d_B] = >_{20} + <_{30} + <_{10} = <_{20}$
$[BsCs] = [d_B] + [BeCs] = <_{10} + >_{13} = <_{3}$
$[BsCe] = [d_B] + [BeCs] + [d_C] = <_{10} + >_{13} + <_{16} = <_{13}$
$[BeCe] = [BeCs] + [d_C] = >_{13} + <_{16} = <_{3}$
$[AsCs] = [AsBs] + [BsCs] = [AsBe] + [BeCs] = <_{27}$
$[AsCe] = [AsBs] + [BsCe] = [AsBe] + [BeCe] = <_{43}$
$[AeCs] = [AeBs] + [BsCs] = [AeBe] + [BeCs] = <_{7}$
$[AeCe] = [AeBs] + [BsCe] = [AeBe] + [BeCe] = <_{23}$

The temporal algebra system is proved as an algebraic group, with associative and transitive relations. We could compute timing from serial specifications. The presentation system schedules media relations and generates SMIL documents automatically after author designing. For example:

Example 2: A presentation is specifying with following media as SMIL document. And the media content is received from content attribute and specifications.

Media A : aff.txt, type:text, freeze
Media B :AChang.gif, type:image, freeze
Media C : http://www.cs.tku.edu.tw/achang/02.gif, type:image
Media D : http://www.mis.kwit.edu.tw/achang/flower.swf, type:12 seconds flash
Media E : GLOBE.avi, type: 10 seconds video
Given temporal constraints:

$As <_3 Bs$; $Bs = Ds$; $Ce <_2 Ds$; $De <_2 Es$;

Value:

Ce=10; Fs=12; Cd=5

The synchronized engine generates a scheduling of consistent document for authors to edit and reference temporal constraints as Figure 2.

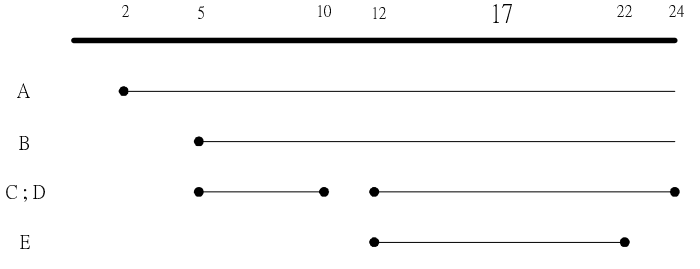


Fig. 2. Scheduling of example 2

The encoder translates the scheduling to a relative SMIL code.

```
<smil >
    <body>
      <par>
            <text id="A" src="aff.txt" begin="2s" fill="freeze"/>
            <img id="B" src="AChang.gif"    begin="5s" fill="freeze" />
            <seq>
                  <img id="C" src=http://www.cs.tku.edu.tw/achang/02.gif begin="5s" end="10s" />
                  <img id="D" src="http://www.mis.kwit.edu.tw/achang/flower.swf"    begin="2s"/>
             </seq>
             <video id="E" src="GLOBE.avi"    begin="12s" />
        </par>
    </body>
</smil>
```

6 Spatial Representation

Let *rs* denote a set of 1-D temporal interval relations. The relation composition table can be refined to a function maps from the Cartesian product of two *rs* to a *rs*. Assuming that f^1 is the mapping function, we can compute f^2, the relation composition function of 2-D objects, and f^3, the one for 3-D objects, from f^1. A conjunction of two 1-D relations denotes a 2-D relation.

Since a 2-D relation is conjunction of two 1-D relations, we use the notation, $rs1 \times rs2$, to denote a 2-D relation set where *rs*1 and *rs*2 are two 1-D relation sets. Thus f^2 is a mapping from Cartesian product of two $rs \times rs$s to a $rs \times rs$. Similarly f^3 is obtained. The following are signatures of these functions:

Functions f^2 and f^3 are computed according to the following formulas:

$$\forall\ i_1 \times j_1,\ i_2 \times j_2 \in \mathrm{P}\ (RelSet \times RelSet)$$

$$f^2\ (i_1 \times j_1,\ i_2 \times j_2) = \prod f^1\ (i_1, i_2) \times f^1\ (j_1, j_2)$$

$$\forall\ i_1 \times j_1 \times k_1,\ i_2 \times j_2 \times k_2 \ \in \mathrm{P}\ (RelSet \times RelSet \times RelSet)$$

$$f^3\ (i_1 \times j_1 \times k_1,\ i_2 \times j_2 \times k_2) = \prod f^1\ (i_1, i_2) \times f^1\ (j_1, j_2) \times f^1\ (k_1, k_2)$$

$$where\ \prod A \times B = \{\ a \times b \mid \forall\ a \in A,\ b \in B\ \}$$

$$\prod A \times B \times C = \{\ a \times b \times c \mid \forall\ a \in A,\ b \in B,\ c \in C\ \}$$

2-D projection representation is the simplest kind of spatial information of practical relevance. Image retrieval algorithms could be constructed based on projection relations. The system is able to handle rotated and reflected images. The matching mechanism focuses on the relative positions among a set of objects instead of searching a single object of a particular shape.

7 Semantic Consistence Checking

In some cases, relation compositions may result in an *inconsistency* due to the user specification or involved events synchronously. For example, if specifications "X before Y with 5 units", "Y before Z with 2 units", and "X after Z" are declared by the user, there exists a conflict between X and Z.

We analyze the domain of temporal relations and use a directed graph to compute the relations of agents. In the computation, we consider all possibilities: the unknown derivations, the multiple derivations, and the conflict derivations. We suggest that the computation domain reveals four types, as discussed below.

The *complete relation domain* is a complete graph which contains possible conflicts. We want to find a *reasonable relation domain* containing no conflict derivation. Note that, in these two domains, both user edges and derived edges exist. If there is a conflict among a set of user edges, one of the user edge must be removed from the cycle, or the relation of that user edge must be re-assigned.

The *reduced relation domain* contains relations specified by the specification only. To avoid the occurrence of conflicts, we place a restriction on the user's interaction. Instead of allowing the user to add an arbitrary relation to the relation graph, we only allow the user to add objects to a *restricted relation domain*, which is a tree and a sub-domain of the reduced relation domain. That is, when the user is about to add a new edge, the user either adds a new node connected to an existing node via an user edge, or joins two sub-trees via the user edge. No cycle is created in the restricted relation domain. Thus, the conflict situation does not exist. The four domains are used in the analysis and computation of object relations.

8 Conclusions

This paper develops an intelligent methodology to discover temporal relationships from multimedia document that supports the presence of multiple distributed multimedia objects, as well as human-computer interaction. It also develops a temporal algebra system as multimedia synchronization model to analyze media presentation time and interaction event. The temporal models are generalized by composing point temporal relations with qualitative and quantitative functions. Temporal semantics can be extracted from the solutions. Authors could easily describe the temporal behavior of multimedia presentation associate hyperlinks with media object over internet. The temporal computation models can be used in many related applications. We believe that the architecture could benefit other interesting researches such as digital archive, multimedia synchronization, temporal data mining and knowledge discovery.

References

1. Allen, J.F.: Maintaining Knowledge about Temporal Intervals. Communications of the ACM 26(11) (1983)
2. Little, T.D.C., Ghafoor, A.: Spatio-Temporal Composition of Distributed Multimedia Objects for Value-Added Networks. IEEE Computer, 42–50 (October 1991)

3. Shih, T.K., Chang, A.Y.: The Algebra of Spatio-Temporal Intervals. In: Proceedings of the 12th International Conference on Information Networking, Japan, January 21-23, pp. 116–121 (1998)
4. Austerberry, D.: The Technology of Video & Audio Streaming, 2nd edn. Focal Press (2005)
5. Drapeau, G.D., Greenfield, H.: Maestro-a Distributed Multimedia authoring environment. In: Proceedings of Summer 1991 USENIX Conference, p. 473 (1991)
6. Vilain, M., Kautz, H.: Constraint Propagation Algorithms for Temporal Reasoning. In: Proceedings AAAI-1986, Philadelphia, pp. 377–382 (1986)
7. Bertino, E., Ferrari, E.: Temporal Synchronization Models for Multimedia Data. IEEE Transactions on Knowledge and Data Engineering 10(4), 612–631 (1998)
8. W3C: Synchronized Multimedia Integration Language (SMIL 3.0). W3C Recommendation (December 01, 2008), http://www.w3.org/TR/2008/REC-SMIL3-20081201/
9. Bulterman, D.C.A.: SMIL 2.0 part 1: overview, concepts, and structure. IEEE Multimedia 8(4), 82–88 (2001)
10. W3C: Timed Text Markup Language (TTML) 1.0. W3C Recommendation (November 18, 2010), http://www.w3.org/TR/2010/REC-ttaf1-dfxp-20101118/

Novel Blind Video Forgery Detection Using Markov Models on Motion Residue

Kesav Kancherla and Srinivas Mukkamala

Department of Computer Science
Institute for Complex Additive Systems and Analysis (ICASA)
Computational Analysis and Network Enterprise Solutuons (CAaNES)
New Mexico Institute of Mining and Technology
Socorro, New Mexico 87801, U.S.A.
{kesav,srinivas}@cs.nmt.edu

Abstract. In this paper we present a novel blind video forgery detection method by applying Markov models to motion in videos. Motion is an important aspect of video forgery detection as it effects forgery detection in videos. Most of the current video forgery detection algorithms do not consider motion in their approach. Motion is usually captured from motion vectors and prediction error frame. However capturing motion for I-frame is computationally expensive, so in this paper we extract the motion information by applying collusion on successive frames. First a base frame is obtained by applying collusion on successive frames and the difference between actual and estimate gives information about motion. Then we apply Markov models on this motion residue and apply pattern recognition on this. We used Support Vector Machines (SVMs) in our experiment. We obtained an accuracy of 87% even for reduced feature set.

Keywords: We would like to encourage you to list your keywords in this section.

1 Introduction

With the recent advances in digital media technologies, availability of low cost, high performance digital cameras and internet, there is an increase in the volume of digital content available. The use of video surveillance cameras for security also added enormous amount of digital data. Due to the digital nature of these files, they can now be manipulated, synthesized and tampered without leaving any visual clues. The availability of sophisticated image/video editing tools further complicated this problem. Thus it is very important to prove the integrity of a digital video.

Forgery detection techniques can be divided into two types: active forgery detection and passive forgery detection. In active forgery detection, a watermark is first embedded into video and any tampering with the video changes the watermark. A similar approach is to extract digital signature from video and detect changes in signature for tamper detection. The major drawback of this approach is the

J.-S. Pan, S.-M. Chen, N.T. Nguyen (Eds.): ACIIDS 2012, Part III, LNAI 7198, pp. 308–315, 2012.

requirement of pre-registration and pre-embedding of video. Most of the current capturing devices do not have an inbuilt watermarking embedding module and users do not prefer watermarks, as they affect the quality of video. Passive methods provide an alternative to this by extracting intrinsic fingerprint from video and apply pattern recognition techniques to perform forgery detection.

Video tampering techniques can be divided into two groups [3]: intra-frame forgery and inter-frame forgery. In intra-frame forgery clippings of object or frames from same video are used to tamper video and in inter-frame forgery clippings of object or frames from different video are used to tamper video. The forgery detection method can blind (forgery attack independent) or forgery attack specific. In this paper we present a blind passive video forgery technique.

In [1] the authors used spatial and temporal artifacts of double Motion Picture Experts Group (MPEG) compression. In an MPEG sequence an I-frame is similar to Joint Pictures Experts Group (JPEG) compression and there is more correlation among frames in a given Group of Pictures (GOP). Thus I-frame double compression is similar to JPEG double compression detection and in a GOP adding a frame or deleting a frame will increase the motion estimation error. Forgery detection of de-interlace and interlaced video is provided in [5]. In their work, they used the motion between fields of a frame and the motion across fields of neighboring frames. However their method is specific to particular video format. In [6] a copy paste forgery is detected by using high correlation between original and forged regions. However their method does not work if copy is taken from different video and high correlation is common in natural videos, thus effecting performance of detection.

Another direction in video forgery detection is the use of noise due to hardware artifacts in camera [3, 4]. In [3] the authors used photon shot noise to detect forged regions. When a video is forged with frames or regions from a different video taken by different camera, the noise characteristics change. The authors used these inconsistencies in noise to detect forged regions. This method works only for inter frame forgery and the authors produced results for static video only. The use of noise residue correlation for detecting forged regions is given in [2]. This work is based on the assumption that tampering regions in video will change the correlation of noise residue between successive frames. The noise residue features provide a reliable forgery detection method; however the detection is effected by quantization noise and motion in video. Even though noise based features provide better accuracy, the performance of such systems is affected by the complex noise extraction process.

In this work we propose a novel video forgery detection scheme which is more computationally efficient when compared to noise based methods. In this work we extract motion based features by modeling motion between frames by Markov models. The motion information is captured by applying an average function on frame window centered at current frame. This method can be used as a coarse detection scheme over fine noise based detection schemes. The rest of the paper is organized as follows: section 2 provides an overview of our approach, in section 3 we provide a brief overview of SVMs, in section 4 we provide the results obtained and in section 5 we conclude our paper.

2 Our Approach

In this work we propose a top-down approach. Where we first detect forged video frames using our technique and latter apply a fine detection method like noise based methods to detect forged blocks. Figure 1 gives the outline of our approach. First the video is converted into set of frames and for each frame we extract the motion information using motion extractor. Later we apply Markov models to this motion residue and extract features for each frame. A pattern recognition algorithm is applied on these features to perform a binary classification. In our experiments we used Support Vector Machines (SVMs).

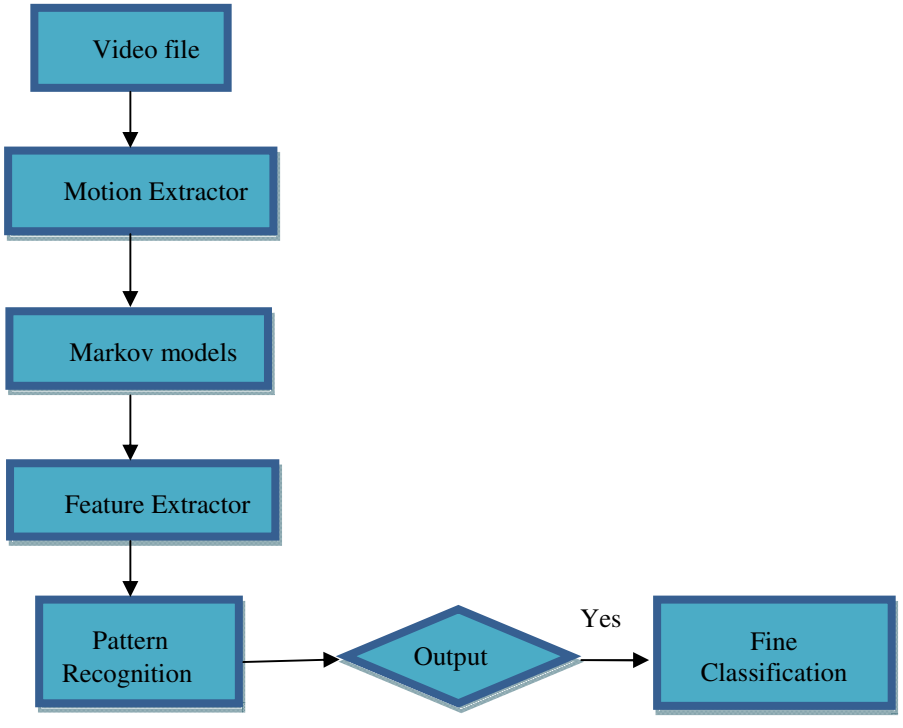


Fig. 1. Basic outline of our approach

The most common way to capture motion in video is to use motion vectors. Motion vectors of P-frame and B-frame can easily be extracted from video bit stream. However motion vector of I-frame is not available readily and extracting it for each frame is computationally expensive. So in our approach we perform a simple average function on frame of window size W which is centered at current frame. Each frame in video can be divided into a base frame and motion between base frame and current frame. Due to temporal redundancy the base frame is common in a given window of frames. So when we average the frame, we extract this base frame. Base frame can be

considered as background in video. Later we subtract the base frame from actual frame we extract motion residue.

After extracting the motion residue, we apply Markov models to it. In order to capture the spatial properties we use two dimensional difference arrays. The difference array is generated by finding the difference between current pixels and neighboring pixels. The difference arrays [17] are calculated along horizontal, vertical, diagonal and minor-diagonal direction. Let motion residue be R (u, v) then the difference arrays along horizontal, vertical, diagonal and minor-diagonal direction are given by equations 1, 2, 3 and 4 respectively.

$$R_h(u,v) = R(u,v) - R(u+1,v) \tag{1}$$

$$R_v(u,v) = R(u,v) - R(u,v+1) \tag{2}$$

$$R_d(u,v) = R(u,v) - R(u+1,v+1) \tag{3}$$

$$R_m(u,v) = R(u+1,v) - R(u,v+1) \tag{4}$$

We model these difference matrix using n-step transition probality matrix. N-step transistion probabilty matrix gives the transistion probabilty between elements that are seperated by n-1 elements. Inorder to reduce to computational complexity, we used only 1-step tranisition probabilty matrix. To further reduce the dimensionality we restrict the value to [-4, 4]. This range is taken by observing the statistical study performed in [17]. Thus the values those are larger than +4 are set to +4 and the values that are smaller than -4 are set to -4. The transition probabilty matrix are given by equations 5, 6, 7 and 8.

$$M_{h(i,j)} = \frac{\sum_{u=1}^{h-2}\sum_{v=1}^{w}\delta(R_h(u,v)=i, R_h(u+1,v)=j)}{\sum_{u=1}^{h-1}\sum_{v=1}^{w}\delta(R_h(u,v)=i)} \tag{5}$$

$$M_{v(i,j)} = \frac{\sum_{u=1}^{h}\sum_{v=1}^{w-2}\delta(R_v(u,v)=i, R_v(u,v+1)=j)}{\sum_{u=1}^{h}\sum_{v=1}^{w-1}\delta(R_v(u,v)=i)} \tag{6}$$

$$M_{d(i,j)} = \frac{\sum_{u=1}^{h-2}\sum_{v=1}^{w-2}\delta(R_d(u,v)=i, R_d(u+1,v+1)=j)}{\sum_{u=1}^{h-1}\sum_{v=1}^{w-1}\delta(R_d(u,v)=i)} \tag{7}$$

$$M_{m(i,\,j)} = \frac{\sum_{u=1}^{h-2}\sum_{v=1}^{w-2}\delta(R_m(u+1,v)=i, R_m(u,v+1)=j)}{\sum_{u=1}^{h-1}\sum_{v=1}^{w-1}\delta(R_m(u,v)=i)} \tag{8}$$

Where h and w are the dimensions of the image and δ = 1 if only if the conditions are satisfied.

3 Experiments and Results

For our experiments, we used 20 different videos downloaded from different forgery dataset sources [8, 9, 10 and11]. This video dataset consists of both copy-paste inpainting forgery video and example-based texture synthesis forgery videos. Datasets [8 and 9] are from NTHU Forensics project, dataset [10] is created by using 'Video Motion Interpolation for Special' and dataset [11] is based on 'Video Inpainting Under Camera Motion'. For more details about forgery dataset creation please refer to their respective sites. First the video dataset is converted into set of frames and for each frame we extract motion residue using our approach. We apply Markov models on this and extract 324 features for each video. To reduce the dimensionality of our feature set we take the average between four sets of features using formula 9.

$$M(i,j) = \frac{M_h(i,j) + M_v(i,j) + M_d(i,j) + M_m(i,j)}{4} \tag{9}$$

The final dataset consists of 5058 frames of which 2786 are from legitimate videos and 2272 are from forged videos. We used SVMs to perform binary classification in our experiment. Details of parameter selection are given in next sub-section. The performance of our forgery detection method is evaluated using accuracy, precision and recall. Table 1 gives the results obtained by using our detection algorithm. The formula for accuracy, precision and recall are given below.

$$Accuracy = \frac{TN + TP}{TN + FP + TP + FN}$$

$$Precision = \frac{TP}{TP + FP}$$

$$\mathrm{Re}\,call = \frac{TP}{TP + FN}$$

Where TN is true negative, TP is true positive, FP is false positive and FN is false negative.

Table 1. Results obtained by using our forgery detection method.

Number of features	Accuracy	Precision	Recall
324	87.09	0.89	0.86
81	87.2	0.89	0.87

3.1 Model Selection

In any predictive learning task, such as classification, both a model and a parameter estimation method should be selected in order to achieve a high level of performance of the learning machine. Recent approaches allow a wide class of models of varying complexity to be chosen. Then the task of learning amounts to selecting the sought-after model of optimal complexity and estimating parameters from training data [14 and 15].

Within the SVMs approach, usually parameters to be chosen are (i) the penalty term C which determines the trade-off between the complexity of the decision function and the number of training examples misclassified; (ii) the mapping function Φ; and (iii) the kernel function such that $K(\mathbf{x}_i, \mathbf{x}_j) = \Phi(\mathbf{x}_i) \cdot \Phi(\mathbf{x}_j)$.

In the case of RBF kernel, the width, which implicitly defines the high dimensional feature space, is the other parameter to be selected [14]. Figure 2 is a SVM model obtained from our experiments, where X-axis is log of cost parameter and Y-axis is log of gamma (kernel) parameter. The different curves are the accuracies obtained for different cost and gamma parameters.

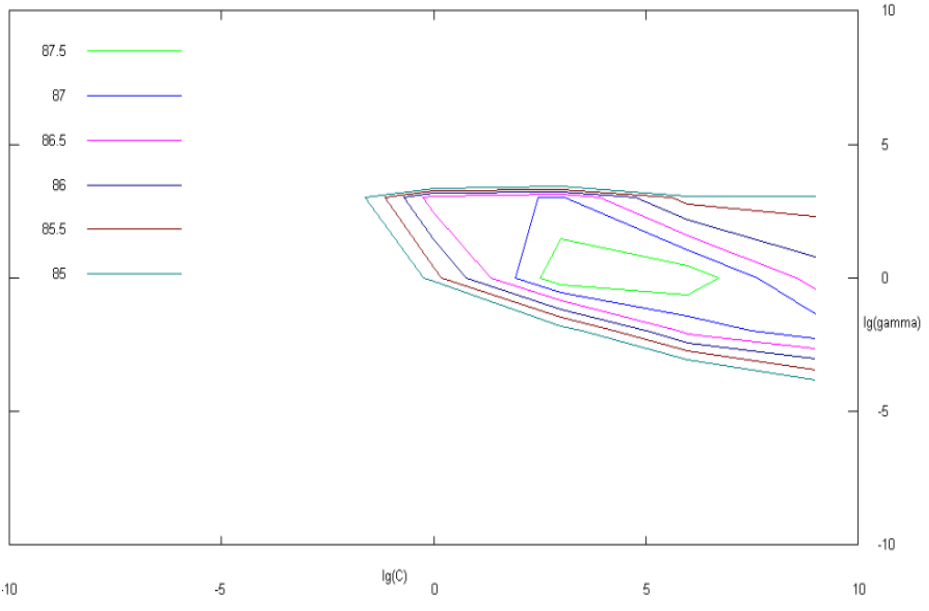


Fig. 2. SVM Model for video forgery detection

3.2 Receiver Operating Characteristic (ROC) Curve

Receiver Operating Characteristic (ROC) Curve is a graphical plot between the sensitivity and specificity. The ROC is used to represent the plotting of the fraction of true positives (TP) versus the fraction of false positives (FP). The point (0, 1) is the perfect classifier, since it classifies all positive cases and negative cases correctly. Thus an ideal system will initiate by identifying all the positive examples and so the curve will rise to (0, 1) immediately, having a zero rate of false positives, and then continue along to (1, 1).

Detection rates and false alarms are evaluated for the phishing data set and the obtained results are used to form the ROC curves. In each of these ROC plots, the x-axis is the false alarm rate, calculated as the percentage of normal video frames considered as steganograms; the y-axis is the detection rate, calculated as the percentage of steganograms detected. A data point in the upper left corner corresponds to optimal high performance, i.e, high detection rate with low false alarm rate [12].The accuracy of the test depends on how well the test classifies the group being tested into 0 or 1. Accuracy is measured by the area under the ROC curve (AUC). An Area of 1 represents a perfect test and an area of .5 represents a worthless test. In our experiment, we got an AUC of 0.9479 using reduced 81 features as shown above in Figure 3.

We obtained an accuracy of about 89% using reduced feature set of 81 features. The major benefit of our approach is computational efficiency over noise based methods. As we only need to perform addition operation to extract motion information and extract features, our method can scale well when compared to other methods. Our method is a blind forgery detection method, so our method can detect unknown forgery attacks. However method cannot provide block level detection but can be used on the top of fine classification algorithm.

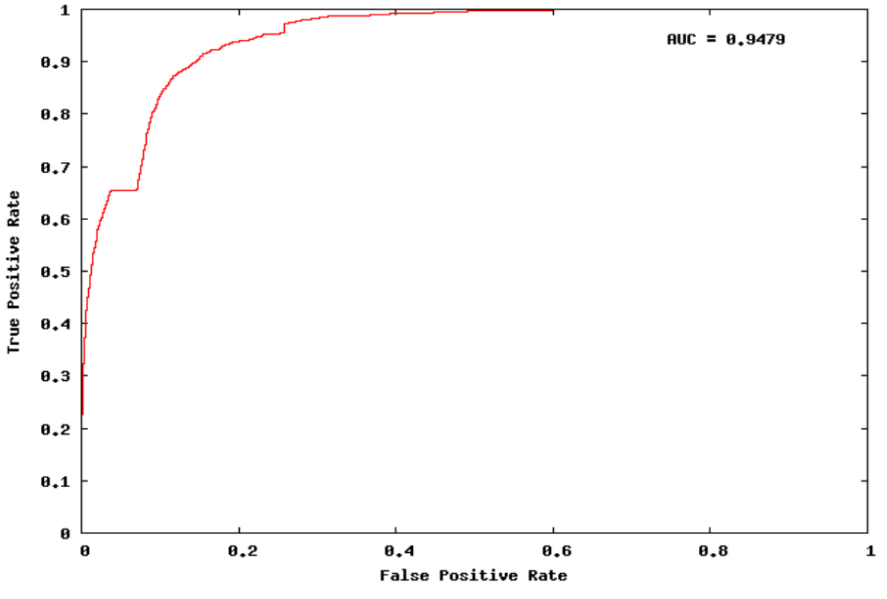


Fig. 3. ROC Curve for reduced feature set

4 Conclusion

In this paper we present a novel blind forgery detection scheme using Markov models of motion frame. First we perform collusion attack on video by applying average filter on window of frames. After that, we extract the motion frame by subtracting actual frame and output of collusion. Then we model this motion frame using Markov models and extract 81 features from each set. Later we reduce the feature vector

dimension by averaging all four sets of features. Experimental results show comparable performance between actual and reduced dataset. As our approach is computationally efficient, this can be used as coarse classification on top of other fine and complex forgery detection algorithms. In the future we would like to perform our experiments on a large dataset. We would also like to perform our feature extraction at bit stream level to further improve scalability.

References

1. Wang, W., Farid, H.: Exposing digital forgeries in video by detecting double quantization. In: Proceedings of the 11th ACM workshop on Multimedia and Security - MM&Sec 2009, New York, NY (2009)
2. Hsu, C., Hung, T., Lin, C., Hsu, C.: Video forgery detection using correlation of noise residue. In: Proceedings of IEEE Workshop Multimedia Signal Processing (MMSP), Cairns, Queensland, Australia, pp. 170–174 (2008)
3. Kobayashi, M., Okabe, T., Sato, Y.: Detecting Forgery From Static-Scene Video Based on Inconsistency in Noise Level Functions. IEEE Transactions on Information Forensics and Security 5(4), 883–892 (2010)
4. Mihcak, M.K., Kozintsev, I., Ramchandran, K.: Spatially adaptive statistical modeling of wavelet image coefficients and its application to denoising. In: Proc. IEEE Int. Conf. Acoustics, Speech, Signal Processing, Phoenix, AZ, vol. 6, pp. 3253–3256 (1999)
5. Wang, W., Farid, H.: Exposing digital forgeries in interlaced and de-Interlaced Video. IEEE Transactions on Information Forensics and Security 2(3), 438–449 (2007)
6. Wang, W., Farid, H.: Exposing digital forgeries in video by detecting duplication. In: Proceedings of the Multimedia and Security Workshop, Dallas, TX, pp. 35–42 (2007)
7. Zhang, J., Su, Y., Zhang, M.: Exposinig digital video forgery by ghost shadow artifact. In: Proc. of ACM Workshop on Multimedia in Forensics, Security and Intelligence, Beijing, China, pp. 49–53 (2009)
8. NTHU Forensics project, `http://www.ee.nthu.edu.tw/cwlin/forensics/forensics.html`
9. NTHU Forensics project, `http://www.ee.nthu.edu.tw/cwlin/inpainting/inpainting.html`
10. Video Motion Interpolation for Special Effect, `http://member.mine.tku.edu.tw/www/TSMC09/`
11. Video Inpainting Under Camera Motion, `http://www.tc.umn.edu/~patw0007/video-inpainting/`
12. Egan, J.P.: Signal detection theory and ROC analysis. Academic Press, New York (1975)
13. Zhao, H., Wu, M., Wang, Z., Liu, K.J.R.: Forensic Analysis of Nonlinear Collusion Attacks for Multimedia Fingerprinting. IEEE Transactions on Image Processing 14(5), 646–661 (2005)
14. Lee, J.H., Lin, C.J.: Automatic model selection for support vector machines, Technical Report, Department of Computer Science and Information Engineering, National Taiwan University (2000)
15. Chang, C.C., Lin, C.J.: LIBSVM: a library for support vector machines, Department of Computer Science and Information Engineering, National Taiwan University (2001)
16. Pevny, T., Fridrich, J.: Merging Markov and DCT features for multi-class JPEG steganalysis. In: Proceedings of SPIE Electronic Imaging, Photonics West, pp. 03–04 (2007)
17. Shi, Y.Q., Chen, C.-H., Chen, W.: A Markov Process Based Approach to Effective Attacking JPEG Steganography. In: Camenisch, J.L., Collberg, C.S., Johnson, N.F., Sallee, P. (eds.) IH 2006. LNCS, vol. 4437, pp. 249–264. Springer, Heidelberg (2007)

Sports Video Classification Using Bag of Words Model

Dinh Duong, Thang Ba Dinh, Tien Dinh, and Duc Duong

Faculty of Information Technology, University of Science
Ho Chi Minh City, Vietnam
dinhseva@gmail.com,
{dbthang,dbtien,daduc}@fit.hcmus.edu.vn

Abstract. We propose a novel approach classify different sports videos given their groups. First, the SURF descriptors in each key frames are extracted. Then they are used to form the visual word vocabulary (codebook) by using K-Means clustering algorithm. After that, the histogram of these visual words are computed and considered as a feature vector. Finally, we use SVM to train each classifier for each category. The classification result of the video is the production of the scores output from all of the key frames. An extensive experiment is performed on a diverse and challenging dataset of 600 sports video clips downloaded from Youtube with a total of more than 6000 minutes in length for 10 different kinds of sports.

Keywords: image classification, sports video classification, bag of words, SVM, K-means, SURF.

1 Introduction

These days, with the rapid growth of technology, video cameras can be purchased with a surprisingly low cost. The consequence of this fact is the tremendous existing amount of videos from both broadcast and personal sources. Automatic video classification becomes a crucial task in video analysis because it is impossible for people to annotate such a vast amount of video resources. In this paper, we focus in classifying sports video shots into different categories, which is important for many applications such as content-based video search, sports strategy analysis, video highlights recommendation. This is a very challenging problem as sports videos are very dynamic and often share similar motion characteristics.

It is worth to emphasize that, most of the current approaches in video classification is inspired by image classification which is still one of the most challenging problems in computer vision. The approaches can be formulated in two categories: discriminative and generative. Discriminative methods often use Bag-of-Words (BoW) representation [6,7], where visual words are local features such as SIFT [1], SURF [2]. Grauman and Darrel also presented the Spatial pyramid matching (SPM) [8] which is efficient for whole image classification. The generative methods, on the other hand, focus on the topic models as in [9].

J.-S. Pan, S.-M. Chen, N.T. Nguyen (Eds.): ACIIDS 2012, Part III, LNAI 7198, pp. 316–325, 2012.

To address the video classification problem, people often choose to narrow down the domain such as: horror movie scenes [10], violent scenes [13]. Some others choose to classify type of camera views in a specific genre of video such as soccer videos [14]. Some approaches try to fuse different cues such as caption, audio, visual information [13, 14].

In this paper, we address the problem of categorizing sports videos due to its popularity and challenges. Takagi [11] focused on camera motion in the video sequence to categorize 6 different sport types. Ling-Yu Duan *et. al.* [12] used top down video shot classification based on pre-defined video shot classes, each of which has a clear semantic meaning. They tested on 4 types of sports and get 85% – 95% of accuracy rate. Recently, Zhang and Guan [15] proposed a large scale video genre classification method using SIFT descriptors in a modified latent Dirichlet allocation (mLDA) framework. The classifier is then built using k-NN algorithm. The method was tested in 23 sports dataset and achieve from 55%-100% accuracy ratios for different categories.

Here we propose a novel method based on Bag-of-Words (BoWs) from which visual words are represented by SURF descriptors. In the experiments, we collected a diverse and challenging dataset with a total of more than 6000 minutes in length. We have tested and analyzed our model intensively on this dataset. The recall and precision analysis shows robust results in all types of sports.

The rest of the paper consists of 3 sections. In section 2, the details of our algorithm are described. Then, Section 3 shows the experiment settings and results followed by the conclusion and future work discussion in Section 4.

2 Our Approach

Our approach is based on bag of visual words model, which originally comes from document representation in terms of form and semantic. The bag of words model has been widely used in classification, recognition, content-based image retrieval and detection [6,7]. Inspired by the image classification algorithm proposed by Csurka et. al. [6], which has been proved to effectively in static image dataset, we propose to classify sports video by extracting key frames, and classify each of them. The final classification result of the video is the production of the scores output from all of the key frames. Our method contains the following 4 main steps:

- Descriptors of video frames are detected and extracted by using SURF approach.
- The descriptors are then used to form up the visual word vocabulary (codebook) by using a cluster algorithm.
- A histogram representation is formed to count the number of visual words appeared in each frame.
- A multi-class classifier considers histogram representation of a frame as a feature vector. It, then, determines which class to assign the test frame to.

Fig. 1 illustrates the 4-step model for sports video classification.

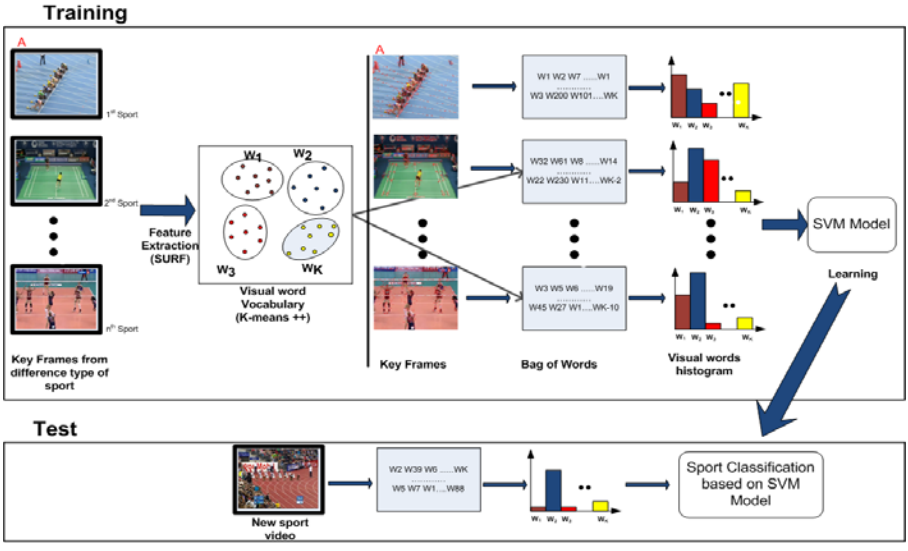


Fig. 1. Illustration of our four steps method base on Bag of Words model

2.1 Descriptors Extraction

Firstly, in the training dataset, key frames are collected base on our pre-defined shot views. Key frames then become input of this step. It then outputs the key points set and their descriptors. Key point is a point in the image, which has a rich local information and is stable under local and global perturbed activities in the image domain, such as affine transformations, scale changes, and rotation. Many key point detectors including Harris, Harris-Laplace, DoG and the descriptors such as SIFT [1], and SURF [2] providing very impressive results. In our framework, we adopt SURF because of its good performance comparing to other methods while having reasonable running time.

Key point detector

SURF detector is based on Hessian Matrix and rely on the determinant of the Hessian for selecting the location and the scale. Given a point $x = (x, y)$ in an image *I*, the Hessian matrix $H(x,\sigma)$ in *x* at scale σ is defined as follows:

$$H(x,\sigma) = \begin{bmatrix} L_{xx}(x,\sigma) & L_{xy}(x,\sigma) \\ L_{xy}(x,\sigma) & L_{yy}(x,\sigma) \end{bmatrix} \quad (1)$$

Where $L_{xx}(x,\sigma)$ is the convolution of the Gaussian second order derivative $\frac{\partial^2}{\partial x^2} g(\sigma)$ with the image *I* in point *x*. It is also similar for $L_{xy}(x,\sigma)$ and $L_{yy}(x,\sigma)$.

The maxima of the determinant of Hessian matrix are then interpolated in scale and image space. Fig. 2 shows an example of the key points detected by SURF.

Fig. 2. Detected key points on tennis, football and swimming

Key point descriptor

The key point descriptor in SURF requires 3 steps. First, it constructs a circular region around the key points and then uses Haar wavelet to compute the orientation in both x and y directions. Finally, SURF descriptors are extracted using square regions. Square regions are split into 4 x 4 sub regions, producing the standard SURF descriptor, SURF-64. Furthermore, another version of SURF descriptor is SURF-128, in which a couple of similar features are added. Because of the advantages of SURF-128 such as more distinctive and not much slower than SURF-64, we apply SURF-128 for key point descriptors in our experimental studies.

2.2 Visual Word Vocabulary Calculation

Assuming that the input contains *n* descriptors, this step aims to partition these *n* descriptors into *k* clusters. Each cluster is called a "visual word". A set of *k* clusters forms the visual vocabulary. After that, the vocabulary is used to represent a video frame, *i.e.* an image. The main idea is that each visual word is represented a region of interest in a frame. Fig. 3 shows how the K-means approach groups similar features in sports frames. To achieve robust performance, a good clustering method and an appropriate codebook size need to be determined. If a codebook size is too small (*i.e.* the value of k is too small), a number of different features could be generally grouped into one cluster. Obviously, this cannot form a good vocabulary and makes the results worse. In contrast, a too large codebook size leads to similar features could be scattered into many clusters making non-sense final results. The codebook size problem is examined carefully in our experiment section. The results are then used to determine a best codebook size for the dataset. Discussions on the relationship between the kind of sports and the codebook size are also mentioned.

Here we use K-means as our clustering method due to its good performance reported in [3]. However, in our empirical experience, when dealing with a large dataset, K-means encounters a number of limitations and the clustering does not give reasonable results. Based on our analysis, we claim that in large datasets, K-means with arbitrary initial cluster-centers cause some blank clusters, which leads to a very bad performance for the system. To avoid this issue, we apply K-means ++ [4] instead. In our experiment results, K-means ++ is proven to be a suitable method for large datasets.

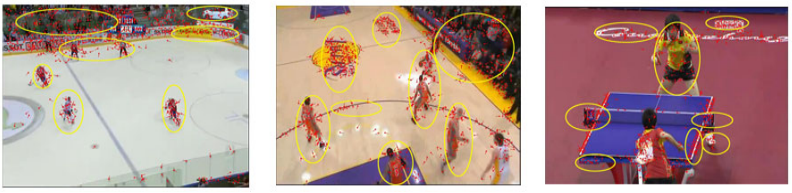


Fig. 3. Similar features are grouped by K – means++

2.3 Histogram Representation

After setting up the visual word vocabulary, in this step, each frame in the training set is reconstructed according to these codebook words. It is then represented by a histogram, which actually counts the number of visual words appearing in the frame (Fig. 1). For instance, in a 100m sprint video, after two steps mentioned above, the features of a frame, named *A*, has been extracted. Assuming that *A* contains n key points, its key point descriptors set is formed as a set of n vectors. Each vector has 128 dimensions based on the SURF-128 descriptor. With each vector, we find its nearest visual word using a distance metric and it means the nearest visual word found happens to appear in *A*. After repeating for all n vectors, we reconstruct *A* as a histogram of visual words appearing in it. Each bin of the histogram is the number of occurrences of a visual word in *A*. For example, in Fig. 1, visual words: 1, 2, 7,…, 3, 200, 101, *k* belong to *A*. In our experiments, three distance metrics: L1, L2 and CHI2 (χ^2) are compared to find the most appropriate metric.

Denote $S_1 = (u_1, u_2, ..., u_m)$ and $S_2 = (w_1, w_2, ..., w_m)$ is the vector representation of two key point descriptors S_1 and S_2

The formula for each distance metric is showed bellow:

$$\text{L1: } D(S_1, S_2) = \sum_{i=1}^{m} |ui - wi| \tag{2}$$

$$\text{L2: } D(S_1, S_2) = \sqrt{\sum_{i=1}^{m} (ui - wi)^2} \tag{3}$$

$$\text{CHI2: } D(S_1, S_2) = \frac{1}{2} \sum_{i=1}^{m} \frac{(u_i - w_i)^2}{u_i + w_i} \tag{4}$$

2.4 Multi-class Classifier

In the last step, we use SVM as a multi–class classifier for our training dataset. The input of this step is a set of <visual word histogram, label> where the label determines the class of a frame. We use one-against-one as approach for multiclass problem where each classification gives one vote to the winning class and the frame is labeled with the class having most votes. In our experiments, we compare three SVM

Kernels: LINEAR, RBF (Radial Basis Function) and POLY (Polynomial) to find the best kernel for our dataset.

3 Experimental Studies and Results

The dataset is a collection of 600 sports videos. The dataset has a total of more than 6000 minutes in length and contains 10 different classes of sports, including 100m sprint, badminton, basketball, fencing, football, hockey, swimming, table tennis, tennis and volleyball. We randomly collected all of these videos from youtube with variation in size and quality. The videos are collected from a large variety of tournament and matches for all sports classes. For example, football videos are collected from UEFA Euro 2012 qualifying, Copa America 2011, Premier League 2010/11, Series A 2010/11, Primera Liga 2010/11, UEFA Champion League 2010/11 and Seagame 24. For each sport, we set up pre-defined shot views based on the camera views. Then, these shot views are used to build the training and testing dataset (Fig. 4). In the first two experiments, we have evaluated the distance metric and codebook size presented in Section 2 with the classification recall and precision reported. The third experiment is aimed to compare the results of three SVM Kernels. The last experiment studies the performance of each kind of sports with the variation of codebook size. Recall and Precision values at each experiment are taken at threshold 0.5.

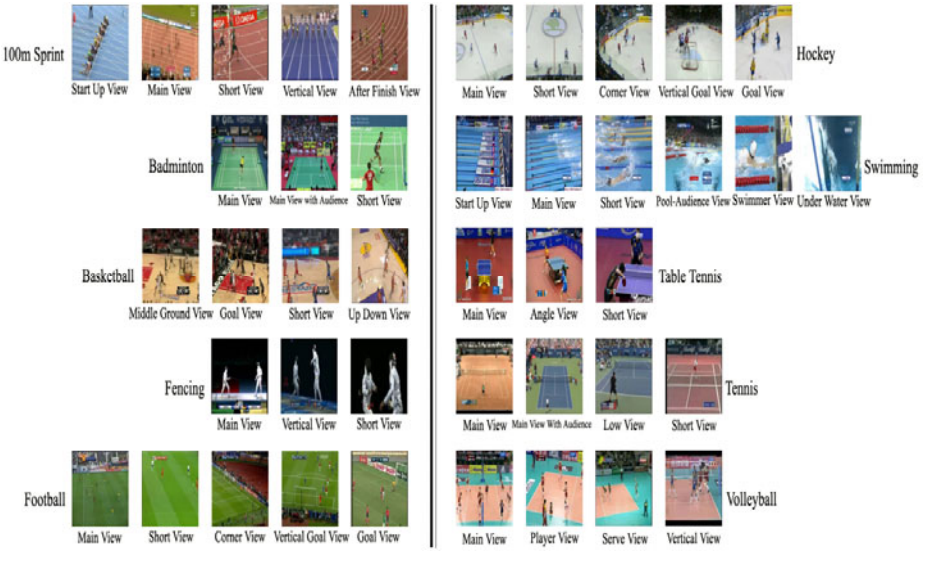


Fig. 4. Pre-defined Shot Views for 100m-Sprint, Badminton, Basketball, Fencing, Football, Hockey, Swimming, Table Tennis, Tennis and Volleyball

3.1 Experiment 1

In the first experiment, we evaluated the performance of three distances: *L1*, *L2* and *CHI2* to the classification. We repeated the experiment with different Codebook sizes.

The SVM-kernel was set to RBF. Tab. 1 reports the average recall and precision values of *10* sport classes for each distance. The results showed that *L2* is the best distance in all cases. The second best is *CHI2* followed by the *L1* distance.

Table 1. Comparing the performance of three distances: L1, L2 and CHI2

Metric	Recall	Precision
L1	92.41%	90.13%
L2	94.04%	96.16%
CHI2	92.84%	90.60%

3.2 Experiment 2

The second experiment is aimed to find the impact of the codebook size to the performance of the classification. In this experiment, *L2* was chosen as the distance metric. The experiment repeated on *9* codebook sizes. In each codebook size, the mean values of three SVM-kernels: Linear, RBF, and Poly are obtained. Tab. 2 shows the recall and precision mean values of 10 sport classes for each codebook size. The results show that the best codebook size is 3850. It is also noticed that 1650 performs a good results. Even a small size of 440 also provides acceptable results. These two codebook sizes are studied in more details in the experiment 4. The running time corresponding to each codebook size is also provided. The last column in Tab. 2 reports the testing of average running time in seconds for each frame. Depending on the application, the best parameters are chosen to compensate the trade-off between the performance and the running time.

Table 2. Comparing the impact of the codebook size

Codebook Size	Recall	Precision	Average time (sec)
110	90.01%	88.86%	0.28
220	92.41%	94.85%	0.41
440	93.48%	95.93%	0.52
880	93.56%	96.29%	0.86
1650	94.32%	97.18%	1.45
2200	94.01%	97.15%	1.86
3300	93.83%	96.93%	2.73
3850	94.28%	97.42%	3.45
4400	94.04%	97.43%	4.88

The last column shows test result on the average running time, taken in a condition of Intel core i3, 2.67ghz and 4Gb of RAM.

3.3 Experiment 3

In this experiment, we compare the results of different SVM-kernels. L2 is the distance metric used. The test procedure is repeated for all of the codebook size. The result is represented by two values indicating the recall and precision. Tab. 3 reports that RBF is always the best kernel for all size. When the dictionary size increases, the performance of LINEAR is improved significantly. This is reasonable according to the [5] experiment conclusion stating that when number of features is large, the data is no needed to map to a higher dimensional space, which means the nonlinear mapping does not improve the performance so much.

Table 3. Comparing the performance of three SVM-kernels: LINEAR, RBF and POLY

Codebook Size	LINEAR		RBF		POLY	
	Re	Pre	Re	Pre	Re	Pre
110	88.8%	87.7%	90.7%	90.0%	90.6%	88.9%
220	91.5%	93.8%	93.1%	95.5%	92.6%	95.2%
440	93.2%	96.0%	93.9%	96.0%	93.3%	95.8%
880	93.2%	96.0%	94.1%	96.7%	93.4%	96.2%
1650	94.3%	97.5%	94.7%	97.3%	94.0%	96.8%
2200	94.1%	97.5%	94.6%	97.3%	93.4%	96.7%
3300	94.2%	97.1%	94.8%	97.3%	92.5%	96.4%
3850	95.1%	97.8%	95.4%	97.8%	92.4%	96.7%
4400	94.9%	97.8%	95.1%	97.8%	92.2%	96.6%

3.4 Experiment 4

Finally, we study the performance of each sport with the variation of codebook sizes. We use *L2* is the distance metric, and RBF is the SVM Kernel. The experiment is repeated with 4 sizes of codebook: 110, 440, 1650 and 3850 (the best code book size). As the result, in Tab. 4, we can observe that badminton, basketball, fencing and table tennis have achieved very high results even if the codebook size is small. When we increase the size, the result does not improve much. This can be explained by the following analysis: these four sports have 2 common characteristics. First, they all contain few numbers of pre-defined shot views. Second, their frames are quite similar with a high rate of repetition. Thus, when represented by the bag of words model, the dictionaries with small size can handle these sports with a high recall and precision values. The best result belongs to Fencing with recall and precision rates are up to 99% and 100%. The poorest is tennis with the recall of 91% and the precision of 96%.

Fig. 6 reports the confusion matrix of 10 sport classes at the best experiment parameters. It shows that some of the most confusing cases are between badminton and tennis, tennis and football, and hockey and volleyball. These confusions are caused by the similarity in sports views and features (shown in Fig. 5). In Fig. 5a, the boundary of the ground and the net is not clear. Thus, a small number of features are taken. This could be confused with short views in football videos. In Fig. 5b, the

confusions could come from the picket along the play ground making the system recognize it as the volley ball net-picket. In Fig. 5c, the shot is very similar to short view of tennis (shown in Fig. 4).

a. Tennis - Football b. Hockey - Volleyball c. Badminton - Tennis

Fig. 5. Common confusion cases

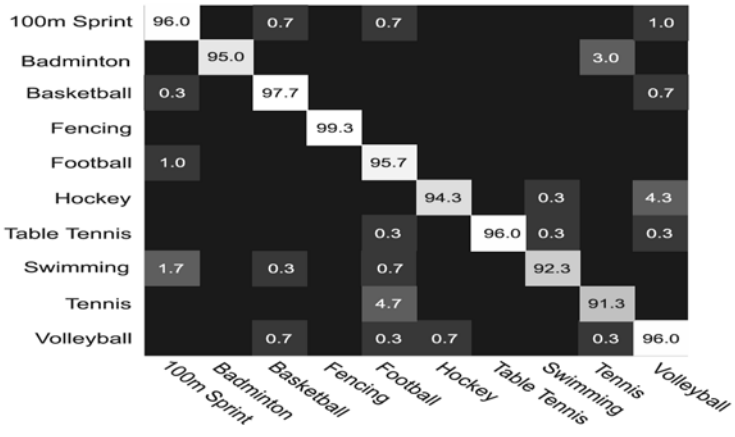


Fig. 6. Confusion matrix of 10 sports at our best experiment parameters

Table 4. Performance of each sport with the variation of codebook size

Code book size	Badminton		Basketball		Fencing		Table Tennis	
	Re	Pre	Re	Pre	Re	Pre	Re	Pre
110	92%	100%	98%	96%	96%	100%	95%	93%
440	94%	100%	98%	97%	99%	100%	97%	99%
1650	94%	100%	98%	97%	99%	100%	96%	100%
3850	95%	100%	98%	98%	99%	100%	96%	100%

4 Conclusions and Future Works

The paper has presented a novel approach to classify sport videos. The proposed method consists of 4 main steps, including descriptors detected and extracted by SURF, visual word vocabulary formed up, the histogram representation constructed and the multi-class classifier used. We have collected a large real-world dataset with a high diversity including 600 videos with a total of more than 6000 minutes for 10

different kinds of sports. As shown in the experiment results, our system shows the Bag of Words model is highly appropriated with sports video shot classification problem. Extensive experiment setups are demonstrated to the advantages of different parameters such as: codebook sizes, classifier kernel. In future, we are going to integrate more sports into the dataset. We will try to improve our model to speed up the running time as well as avoid confusions. We also would like to integrate with state-of-the-art shot boundary detection to automatically provide the shots for classification.

References

1. Lowe, D.G.: Distinctive image features from scale-invariant keypoints. International Journal of Computer Vision 60, 91–110 (2004)
2. Bay, H., Ess, A., Tuytelaars, T., Gool, L.V.: Speeded-Up Robust Features (SURF). Computer Vision and Image Understanding 110(3), 346–359
3. Bosch, A., Zisserman, A., Muñoz, X.: Scene Classification via pLSA. In: Leonardis, A., Bischof, H., Pinz, A. (eds.) ECCV 2006. LNCS, vol. 3954, pp. 517–530. Springer, Heidelberg (2006)
4. Arthur, D., Vassilvitskii, S.: k-means++: the advantages of careful seeding. In: Proceedings of the Eighteenth Annual ACM-SIAM Symposium on Discrete Algorithms, pp. 1027–1035 (2007)
5. Hsu, C.-W., Chang, C.-C., Lin, C.-J.: A Practical Guide To Support Vector Classification. Technical Report, Department of Computer Science and Information Engineering, National Taiwan University, Taipei (2003)
6. Csurka, G., Dance, C.R., Fan, L., Willamowski, J., Bray, C.: Visual Categorization with Bags of Keypoints. In: ECCV 2004 (2004)
7. Sivic, J., Russell, B.C., Efros, A.A., Zisserman, A., Freeman, W.T.: Discovering object categories in image collections. In: ICCV 2005 (2005)
8. Grauman, K., Darrel, T.: The pyramid match kernel: discriminative classification with sets of image features. In: ICCV 2005 (2005)
9. Fei-Fei, L., Perona, P.: A Bayesian Hierarchical Model for Learning Natural Scene Categories. In: CVPR 2005 (2005)
10. Moncrieff, S., Venkatesh, S., Dorai, C.: Horror film genre typing and scene labeling via audio analysis. In: ICME 2003 (2003)
11. Takagi, S., Hattori, S., Yokoyama, K., Kodate, A., Tominaga, H.: Sports Video Categorizing Method Using Camera Motion Parameters. In: 2003 International Conference on Multimedia and Expo. (ICME 2003), vol. 2 (2003)
12. Duan, L.-Y., Xu, M., Tian, Q.: Semantic Shot Classification in Sports Video. In: Proc. of SPIE Storage and Retrieval for Media Database 2003, pp. 300–313 (2003)
13. Nam, J., Alghoniemy, M., Tewfik, A.H.: Audio-visual content-based violent scene characterization. In: ICIP 1988 (1998)
14. Lang, C., Xu, D., Jiang, Y.: Shot Type Classification in Sports Video Based on Visual Attention. In: ICCINC 2009 (2009)
15. Zhang, N., Guan, L.: An efficient framework on large-scale video genre classification. In: MMSP 2010 (2010)

Self Health Diagnosis System with Korean Traditional Medicine Using Fuzzy ART and Fuzzy Inference Rules

Kwang-Baek Kim[1] and Jin-Whan Kim[2,⋆]

[1] Dept. of Computer Engineering, Silla University, Korea
[2] Dept. of Computer Engineering, Youngsan University, Korea

Abstract. Korean traditional medicine has obtained more attention from the public and IT service industry especially after 'Dong-eui-bo-gam' was registered to UNESCO Memory of the World. However, there are many obstacles in developing and commercializing an on-line self-diagnosis system by Korean traditional medicine. From the service point of view, since people are accustomed to the westernized style of diagnosis (symptom-disease pair), it is not easy to understand what traditional Korean medicine diagnoses and how one can react. Technically speaking, we need a special symptom-disease database because Korean traditional medicine has been built upon the innate characteristics of Korean people's body. Thus, in this paper, we propose a self-diagnosis system of Korean traditional medicine based on Korean Standard Causes of Death Disease Classification Index (KCD) and fuzzy ART/inference method. Since this is for self-diagnosis, our system has graphical user-friendly interface that accepts symptoms of user from a certain part of body where the user feels inconvenient. Then, fuzzy ART algorithm and fuzzy inference engine picks up five most probable diseases with their causes and treatments extracted from Korean traditional medicine books. The power of our system comes from a fuzzy inference module combined with fuzzy ART algorithm that helps classifying related disease from database with accuracy. Our system is verified by field experts of Korean traditional medicine in collecting symptom-disease-treatments relationships and performance evaluation of experiment results.

Keywords: self health diagnosis, Korean traditional medicine, fuzzy ART.

1 Introduction

Korean traditional medicine has been frequently recognized as magical but unscientific folk remedy. However, the legendary textbook of Korean traditional medicine, 'Dong-eui-bo-gam', was registered to UNESCO Memory of the world in 2009 after year-long verification processes by International Advisory committee of UNESCO. That means Korean traditional medicine is internationally

⋆ Corresponding author.

J.-S. Pan, S.-M. Chen, N.T. Nguyen (Eds.): ACIIDS 2012, Part III, LNAI 7198, pp. 326–335, 2012.

recognized as a sufficiently reasonable and scientific medical treatment as western medicine.

However, diagnosis based on Korean traditional medicine is not easy to understand by the public as its inference mechanism is largely metaphorical or abstract. While there are plenty of internet services for westernized medicine for self-diagnosis, causes and treatment information [1] [2], there are few such services for Korean traditional medicine. Furthermore, those rare services can only give information of treatments and symptoms with given name of disease. Thus, it is post-hoc supplementary information after expert's diagnosis.

The main difficulties in building informative pre-diagnosis or self-diagnosis system is that it requires a reliable symptom-disease database and disease classification system. For the symptom-disease database, since Korean traditional medicine has been built upon the innate characteristics of Korean people's body [3] [4] [5] [6], we need such specialized evidences for Koreans and happily the government (Statics Korea, http://www.kostat.go.kr) has published Korean Standard Causes of Death Disease Classification Index (KCD) that can be a perfect starting index of such Korean specialized database build-up.

For the disease classification system, we need a unsupervised learning method since a supervised learner such as neural network algorithm causes frequent relearning as symptoms and diseases are to be added. Thus in this paper, we propose a self-diagnosis system for Korean traditional medicine with a disease database based on KCD and a fuzzy ART algorithm in conjunction with fuzzy inference engine as a classification system. Previous studies [7] [8] [9] [10] [11] have explored several neural/fuzzy unsupervised learning algorithms and found that fuzzy ART is most appropriate with its stability and accuracy in diagnosis.

Given symptoms from the user, guided by graphical user interface, our system extracts five most probable diseases with causes and treatments. In those processes, fuzzy inference algorithm plays a critical role to pick up target diseases accurately. The performance of our system as well as relationships among symptoms and diseases and their causes/treatments are verified by experts in Korean traditional medicine and the result is sufficiently competitive as a self-diagnosis system.

The structure of this paper is as following; Section 2 describes the process of collecting symptom/disease data and the structure of disease database entity. Section 3 explains disease classification systems and its main algorithms - fuzzy ART and fuzzy inference rules. Then section 4 shows the result and analysis of experiments, followed by concluding remarks.

2 Preparing Database

We collect 739 diseases and 363 related symptoms based on KCD (Korean Standard Causes of Death Disease Classification Index) which replaces diseases of ICD (International Causes of Death) published by WHO with Korean traditional medicine. Information on causes and treatments of such diseases are extracted from many textbooks of Korean traditional medicine including well

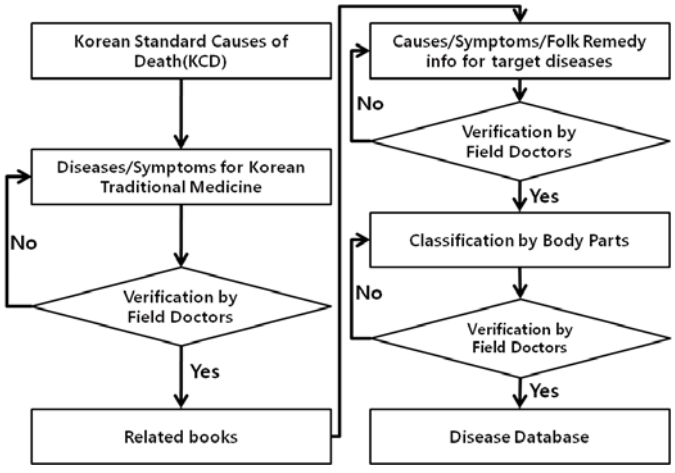


Fig. 1. Process of constructing disease database collection and symptoms with respect to related body part

Table 1. Entity definitions

Entity	Explanation
disease	Disease code, ICD code, KCD code, Disease names (Korean medicine and Western Medicine), Causes, Treatments
symptom	Symptom code, Symptom name, related body part
bodypart	Body part code, Part name
guest	User information
doctor	Expert/medical doctor's information

known folk remedies. Collected information is verified by doctors in Korean traditional medicine three times to reduce conflicts and errors. Collected symptoms are classified into 17 groups with respect to related human body parts. Figure 1 summarizes database construction processes.

Database entities are designed as shown in Table 1. It includes user information and disease information. Disease information includes ICD code and KCD code and names of diseases in Korean and western medicine so that user can get maximal information about the disease.

3 Disease Classification System with Fuzzy ART and Fuzzy Inference Algorithm

In general, neural network learning algorithms have common inefficiency in that they require frequent relearning steps as diseases/symptoms are added/deleted/updated since they are supervised learners. Thus, we need a unsupervised learner that has maximal independencies on clusters with respect to input. Previous

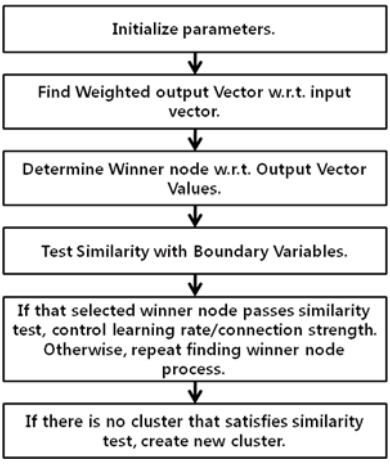


Fig. 2. Processes of fuzzy ART algorithm in disease classification

studies [7] [8] [9] [11] tested various forms of FCM (Fuzzy C-Means) and ART algorithms in diagnosis and verified that an enhanced fuzzy ART is the best in stability.

In that algorithm [11], with given user input (most representative symptoms), it sends queries to database to suppress selecting unrelated symptoms. However, for some diseases that have different symptoms in their early and latter stage, or disease like flu which has too many symptoms but some are inherent to user, this algorithm has a certain level of weakness because in those cases, by the characteristics of the fuzzy ART that applies average of input patterns in controlling weights, the similarity between disease vector and symptom vector becomes unexpectedly low. Thus, we add a fuzzy inference engine to mitigate that weakness. The overall process of disease classification system is described as Fig. 2 and explanations at each step will be followed.

In step 2, the weight is to set the maximum of output vector as 1. There are cases that the maximum of output vector is less than 1 due to the inconsistency between input binary vector and connecting weight vector that has continuous nature. This weight is applied to the disease classification and membership function of disease with respect to symptoms in the learning process.

In step 3, there are two different cases for computing output vector. The first case is for the early learning process. The output vector is the minimum of fuzzy membership value divided by output weight and boundary variable value divided by output weight. The winning neuron is determined by taking the maximum of such output vector values.

In that case, the process continues to step 4 to step 6 such that after investigating the similarity between boundary variables, it determines if a new cluster is required. If it is sufficiently similar (step 5), it controls learning rate and connection strengths. Otherwise, it creates a new cluster (step 6). The final output

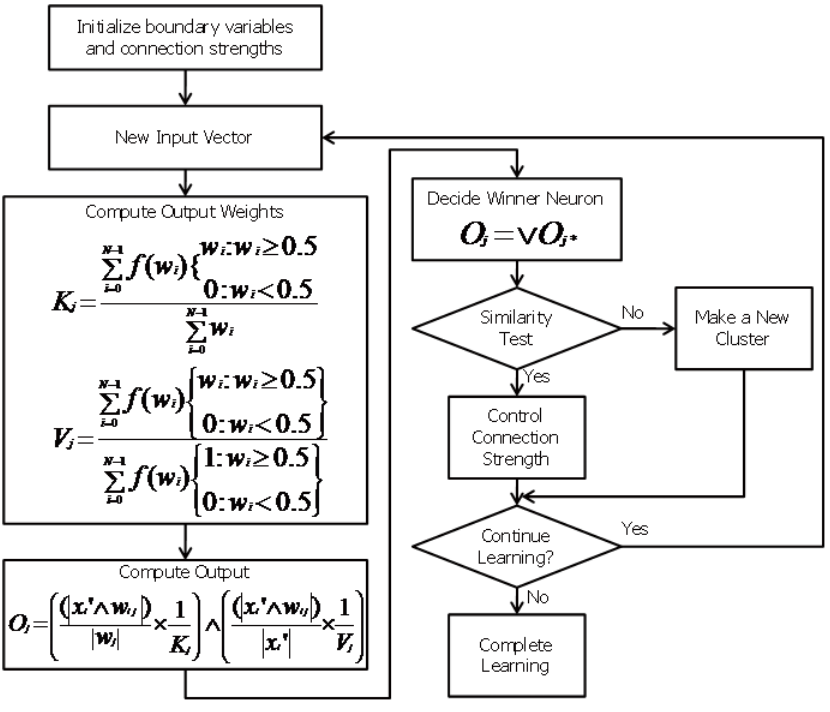


Fig. 3. Flow chart of the first case

vector means the degree of membership belongs to the disease. The flow chart of such process is shown in Fig. 3 [9].

The second case is when a user wants to know the diagnosis with input. The criteria to determine five most probable diseases are related to the agreement rate between user symptoms and disease symptoms (variable X) and the number of symptoms belongs to the disease (variable Y). Values of two variables are computed by related membership functions and then fuzzy inference rules are applied to finalize the result. Then, it is multiplied to the value of the first case to control the degree of membership with respect to user input.

$$Oj = \sum_{i}^{n-1} Output[i] * Ot \tag{1}$$

In Eq. (1), *Output* is the output of the first case and *Ot* is the result of the second case after applying fuzzy inference rules. If *Ot* is high, *Output* also becomes high because high *Ot* means that it is highly related to the user input symptoms and if *Ot* is low, by the same reason, *Output* also becomes low. *Ot* is computed as Eq. (2).

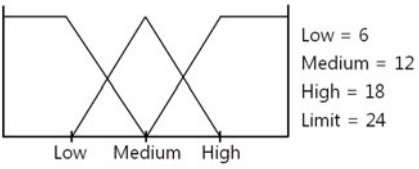


Fig. 4. Membership function for the agreement rate between user symptoms and disease symptoms

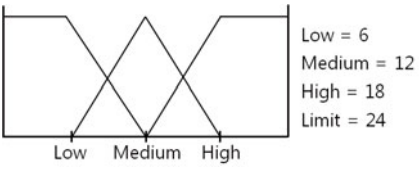


Fig. 5. Membership functions for the number of symptoms belongs to the disease

$$Ot = Fuzzy\left(\begin{array}{l}\#\text{ of symptoms according with user symptoms,}\\ \#\text{ of symptoms of all diseases}\end{array}\right) \tag{2}$$

There are three fuzzy intervals with respect to the agreement rate of symptoms between user symptoms and disease symptoms, L (low), M (medium), H (high) and its membership function is shown as Fig. 4.
(1) When the agreement rate between user symptoms and disease symptoms is low:

$$\begin{array}{l} if(X \leq low)\ then\ \mu(X) = 1 \\ else\ if(X \geq medium)\ then\ \mu(X) = 0 \\ else\ \mu(X) = \frac{medium - X}{medium - low} \end{array}$$

(2) When the agreement rate between user symptoms and disease symptoms is medium:

$$\begin{array}{l} if(X \leq low)\ or\ (X \geq high)\ then\ \mu(X) = 1 \\ else\ if(X \geq medium)\ then\ \mu(X) = \frac{high - X}{high - medium} \\ else\ if\ (X \leq medium)\ then\ \mu(X) = \frac{X - low}{medium - low} \end{array}$$

(3) When the agreement rate between user symptoms and disease symptoms is high:

$$\begin{array}{l} if(X \geq high)\ then\ \mu(X) = 1 \\ else\ if(X \leq medium)\ then\ \mu(X) = 0 \\ else\ \mu(X) = \frac{X - medium}{high - medium} \end{array}$$

There are three fuzzy intervals with respect to the total number of symptoms belong to the disease, L (low), M (medium), H (high) and its membership function is shown as Fig. 5.

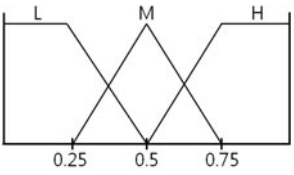


Fig. 6. Disease membership functions for the user input

R1 : If X is L, Y is L then α is L
R2 : If X is L, Y is M then α is M
R3 : If X is L, Y is H then α is M
R4 : If X is M, Y is L then α is L
R5 : If X is M, Y is M then α is M
R6 : If X is M, Y is H then α is H
R7 : If X is H, Y is L then α is M
R8 : If X is H, Y is M then α is H
R9 : If X is H, Y is H then α is H

Fig. 7. Fuzzy inference rules

(1) When the total number of symptoms belongs to the disease is low:

$$if(Y \leq low)\ then\ \mu(Y) = 1$$
$$else\ if(Y \geq medium)\ then\ \mu(Y) = 0$$
$$else\ \mu(Y) = \frac{medium - Y}{medium - low}$$

(2) When the total number of symptoms belongs to the disease is medium:

$$if(Y \leq low)\ or\ (Y \geq high)\ then\ \mu(Y) = 1$$
$$else\ if(Y \geq medium)\ then\ \mu(Y) = \frac{high - Y}{high - medium}$$
$$else\ if\ (Y \leq medium)\ then\ \mu(Y) = \frac{Y - low}{medium - low}$$

(3) When the total number of symptoms belongs to the disease is high:

$$if(Y \geq high)\ then\ \mu(Y) = 1$$
$$else\ if(Y \leq medium)\ then\ \mu(Y) = 0$$
$$else\ \mu(Y) = \frac{Y - medium}{high - medium}$$

Then the disease membership function Ot is shown as Fig. 6.

The $if - then$ style fuzzy inference rules (Figure 7) to determine most probable disease based on user input are as following by combining above two fuzzy membership functions. We apply Min-Max inference method in computing membership degree of diseases. The overall flow of the second case can be summarized as Fig. 8.

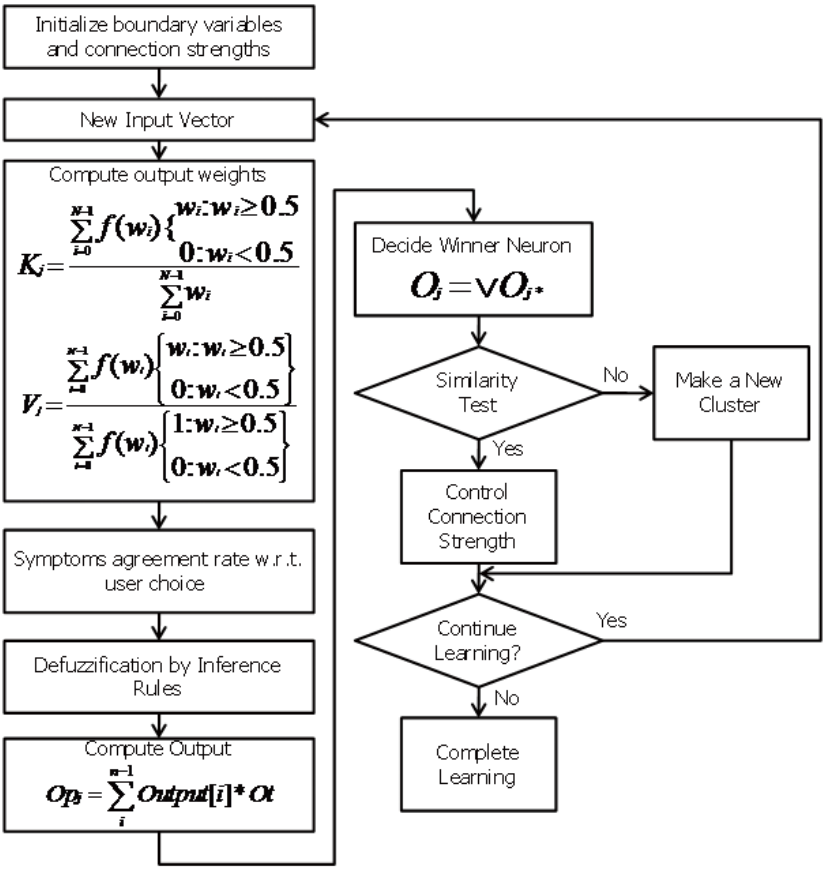


Fig. 8. Flow chart of the second case

4 Experimental Results

The implementation environment is as following; a IBM compatible PC with Intel Pentium IV 2 GHz CPU and 1G RAM is used and Eclipse 3.2, Apache Toolcat 5.5, Apache 2.2, Adobe Photoshop 7.0, JDK 1.6 and Oracle 10g are used in implementation and the system is available for on-line environment using JSP. Figure 9 shows the block diagram of the proposed system.

Users are guided by the graphical interface to give input symptoms. Symptoms are classified into 17 representative body parts and the user clicks the body part from the human body image to select the input symptoms. Then, the system outputs five most probable diseases with their causes and treatments by fuzzy ART algorithm and fuzzy inference rules explained in section 3.

This result can be compared with the result of previous study [9] that uses fuzzy ART but no fuzzy inference rules. In order to have fair comparison, we give the same input - cephalalgia (headache), fatigue, and anorexia to systems with

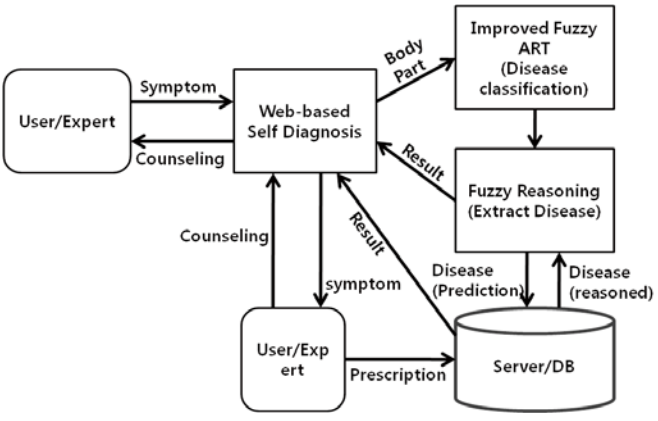


Fig. 9. Block diagram of the proposed self-diagnosis system

Table 2. Result of fuzzy ART algorithm (without inference rules)

Disease (Korean/Western)	Agreement
Kang-Hwa-Sang-Yom (Conjunctivitis)	2/4
Gan-Huh-Jung (Viral Hepatitis)	0/12
Pung-Han-Hyul-Tong (Colesystitis)	0/7
Dam-Huh-Jung (Hydrops of Galbaladder)	0/9
No-kwon-Sang (Neurastenia)	2/8

Table 3. Result of fuzzy ART algorithm (with inference rules)

Disease (Korean/Western)	Agreement
Kang-Wha-Sang-Yom (Conjunctivitis)	3/8
Sang-Cho-Wha	3/8
Migraine	3/11
Infectious	3/14
Yung-Hyul-Huh-Son (Multiple Mayeloma)	3/16

and without fuzzy inference rules. When fuzzy inference rules are not applied, five most probable diseases are extracted and the number of agreed symptoms for those diseases is shown as Table 2. When fuzzy inference rules are applied, five most probable diseases are extracted and the number of agreed symptoms for those diseases is shown as Table 3. Names of diseases are represented by Korean medicine first and then corresponding western medicine disease name in parenthesis. This analysis is done by Korean traditional medical doctor. As one can see the comparison in Table 3, fuzzy inference rules play a critical role in our system to have a more accurate diagnosis.

5 Conclusion

From the previous exploration of several fuzzy unsupervised learners in diagnosis, we select an enhanced fuzzy ART algorithm as the main engine of disease classification system. However, with only fuzzy ART, it is not satisfactory in accuracy. Thus, we applied fuzzy inference rules based on the agreement rate between user symptoms and disease symptoms and the number of symptoms belongs to the disease. The fuzzy inference engine gives the proposed system high enough accuracy to be used as an auxiliary tool for Korean traditional medical doctors in diagnosis or as a self-diagnosis system.

This system gives the users instant access to information about the target diseases with causes and treatments from given user's symptoms. It will be available to the users in the shape of smartphone applications (app) as well as web applications. We partner with app evangelist to implement a successful shift to sustainable diagnosis system. Thus, this system is ready to be used in real world applications.

Acknowledgement. This paper was supported by the research funding of Youngsan University.

References

1. http://www.healthkorea.net
2. http://www.burimhong.pe.kr
3. Kim, Y.S.: Dong-eui-bo-gam. Solbit Publications (2003)
4. N.T.T.C. for Health Personnel, Medical Treatments for Family Doctors. Seoul National University (1993)
5. Lee, S.J.: Dong-eui-bo-gam by Symptoms. Obi Enterprise (2004)
6. Lee, C.H.: Chinese Medicine. Minjoong Seogwan (1999)
7. Han, K.T., Kim, H.C., Ko, J.Y., Lee, C.W.: Real-time intrusion detection using fuzzy adaptive resonance theory. The Journal of Korean Institute of Information Scientists and Engineers 28(2), 640–642 (2001)
8. Kim, K.B., Kim, M.N., Woo, Y.W., Noh, H.C., Shin, S.H.: Self health diagnosis system of oriental medicine using fuzzy inference method. In: Proceeding of Winter Conference, pp. 207–211. Korea Society of Computer Information (2010)
9. Kim, K.B., Kim, M.N., Cho, S.K., Noh, H.C.: Self health diagnosis system of oriental medicine using enhanced fuzzy ART algorithm. In: Proceeding of Summer Conference, pp. 329–332. Korea Society of Computer Information (2009)
10. Kim, K.-B., Kim, S., Sim, K.-B.: Nucleus Classification and Recognition of Uterine Cervical Pap-Smears Using Fuzzy ART Algorithm. In: Wang, T.-D., Li, X., Chen, S.-H., Wang, X., Abbass, H.A., Iba, H., Chen, G.-L., Yao, X. (eds.) SEAL 2006. LNCS, vol. 4247, pp. 560–567. Springer, Heidelberg (2006)
11. Kim, K.B., Woo, Y.W., Kim, J.S.: Self disease diagnosis system using enhanced ART2 algorithm. The Journal of The Korea Institute of Maritime Information and Communication Sciences 11(11), 2150–2157 (2007)

Real-Time Remote ECG Signal Monitor and Emergency Warning/Positioning System on Cellular Phone

Shih-Hao Liou[1], Yi-Heng Wu[1], Yi-Shun Syu[1], Yi-Lan Gong[1], Hung-Chin Chen[2], and Shing-Tai Pan[1,*]

[1] Department of Computer Science and Information Engineering, National University of Kaohsiung, No. 700, Kaohsiung University Rd., Nanzih Dist., Kaohsiung 811, Taiwan, R.O.C.
stpan@nuk.edu.tw
[2] Institude of Computer Science and Information Engineering, National University of Kaohsiung, No. 700, Kaohsiung University Rd., Nanzih Dist., Kaohsiung 811, Taiwan, R.O.C.

Abstract. This paper implements a remote real-time health care system mainly based on electrocardiogram (ECG), body temperature, pulse-based real-time monitoring. A cellular phone with Android O.S. and global positioning system (GPS) is adopted as the platform for this system. The monitor of electrocardiogram (ECG) is performed by a statistical model, Hidden Markov model (HMM), to immediately determine the status of the patient's body. Besides, an automatic warning and positioning system is designed so that the patients can receive timely rescue. Also, a suggestion, if necessary, for finding the closest hospital will be given by this system. In this system, a device for measuring ECG signal is attached on a patient's body and remotely transfers the ECG data to cellular phone through Bluetooth device. The ECG data are then transferred to and stored in the server through internet. All the data in the sever for a patient are used to train and update the HMM model in the cellular phone to get a more precise prediction of the patient's health. Experiments in this paper show that the implemented system works well and is helpful to people's health care.

Keywords: ECG, HMM, Android, Cellular Phone.

1 Introduction

Health care monitor by computer is increasingly important for people with the development of IT technology. Most of the medical equipment are inconvenient to carry in hospital. People want to do the physical checkup might go to the hospital. If an emergency disease happened like heart disease, people always can't handle it. So we want to prevent the case happening. In this paper, we monitor the ECG signal through a cellular phone with Android system.

* Corresponding author.

J.-S. Pan, S.-M. Chen, N.T. Nguyen (Eds.): ACIIDS 2012, Part III, LNAI 7198, pp. 336–345, 2012.

An electrocardiogram (ECG) signal is the manifestation of the myocardium electrical activity on the body surface, which appears as a nearly periodic signal [1]. Traditionally, the ECG cycle is labeled using the letters P, Q, R, S, and T for the individual peaks of the whole cycle's waveform. There are several investigations dealing with the analysis of ECG signals for cardiac arrhythmia. For instance, some of the most popular descriptors which are based on assessment of the QRS complex morphology using pattern recognition methods have been proposed [2, 3]. The papers [4-6] proposed some signal detection methods for discriminate cardiac arrhythmia in the time or frequency domain. Different transforms such as Hilbert transform [7], cross-distance analysis [9], wavelet transform [10], and Hermite function [11] are used to automatically detect, classify and analyze ECG beat. In this study, Hidden Markov Model (HMM) is used to analyze the ECG signal.

The Hidden Markov Model is a powerful statistical tool for modeling generative sequences that can be characterized by an underlying process generating an observable sequence. Andrei Markov gave his name to the mathematical theory of Markov processes in the early twentieth century, but it was Baum and his colleagues that developed the theory of HMMs in the 1960s. HMMs have found applications in many areas in signal processing, particularly in speech recognition. It had also been applied with success to low level NLP tasks such as part-of-speech tagging, phrase chunking, and extracting target information from documents. However, HMM isn't applied to ECG recognition yet. It is the purpose of this paper to implement the HMM-based ECG monitor on a cellular phone with Android system. With the implemented system, we can monitor remotely the heart beats of some people.

2 Hidden Markov Model

The HMM allows for the analysis of non-stationary multivariate time series by modeling both the state transition probabilities and the probability of observation of a state. In the HMM process, the result of the previous state will influence the state recognition result of the next state. This is similar to the process for heart staging that should consider the relationship between the previous heart stage and the next heart stage. Hence, the HMM is a promising model for heart staging which possesses the properties of successive stage transition and a novel strategy based on the HMM for ECG analysis is proposed in this paper. According to the type of the probability distributions used in HMMs, HMMS can be categorized as Continuous Hidden Markov Model (CHMM) and Discrete Hidden Markov Model (DHMM). The DHMM provides more stable recognition results and faster training with the recognition accuracy that is no less than CHMM. Therefore, the DHMM is adopted for ECG analysis in this paper. The useful features of the ECG signals are selected to training the DHMM. In order to rule out the impossible or rare situations of stage transition, the probability matrices in the proposed transition-constrained DHMM were also adjusted to accommodate the transition of heart stages in the training phase to improve the recognition rate. The constructed HMM model can become a reliable computer-assisted tool for the clinical staff to increase the efficiency of heart recognition in the future.

2.1 Train DHMM

Figure 1 illustrates the training process of DHMM model. First, the matrix A, B and π which describe the DHMM model and will be explained in the following text, are randomized at initial step. Then, the speech features are quantized through the trained codebook. The quantized features are then the observation of the DHMM model. The corresponding probability of the observation can be found from the present value of A, B and π. Using these probabilities, we can run the Viterbi Algorithm [13] to update the matrix A, B and π until the values in these matrix converge.

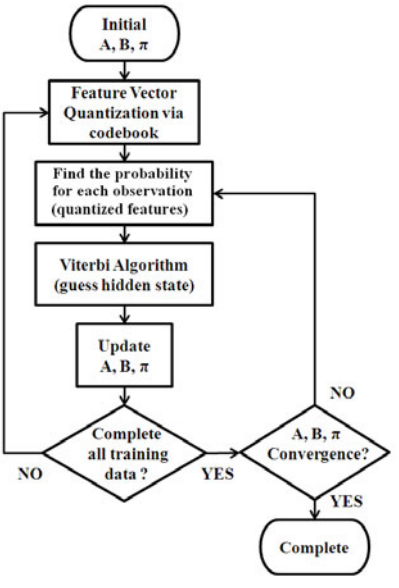


Fig. 1. The DHMM model training process

In a DHMM, the hidden states are always unobservable while the outputs in each state are observable. Each hidden state has a probability distribution over the possible output tokens (i.e., the observation). Therefore, the sequence of output tokens generated by the DHMM gives some information about the sequence of states. For the purpose of clarification, the relation between the features of speech, the observation and the hidden states of DHMM is depicted in Fig. 2. In the following, the definition and detail training method of the parameters in DHMM are introduced [13]. First, the definition of parameters in DHMM is introduced as follows.

λ: DHMM model, $\lambda = (A, B, \pi)$

A: $A = [a_{ij}]$, a_{ij} is the probability of state x_i transferring to state x_j, $a_{ij} = P(q_t = x_j \mid q_{t-1} = x_i)$

B: $B = [b_j(k)]$, $b_j(k)$ is the probability of kth observation which is observed from the state x_j, i.e., $b_j(k) = P(o_t = v_k \mid q_t = x_j)$

π: $\pi = [\pi_i]$, π_i is the probability of the case where the initial state is x_i, $\pi_i = P(q_1 = x_i)$

X: the state vectors of DHMM, $X = (x_1, x_2, \cdots, x_N)$

V: the observation event vector of DHMM, $V = (v_1, v_2, \cdots, v_M)$

O: the observation results of DHMM, $O = o_1, o_2, \cdots, o_T$

Q: the resulting states of DHMM, $Q = q_1, q_2, \cdots, q_T$

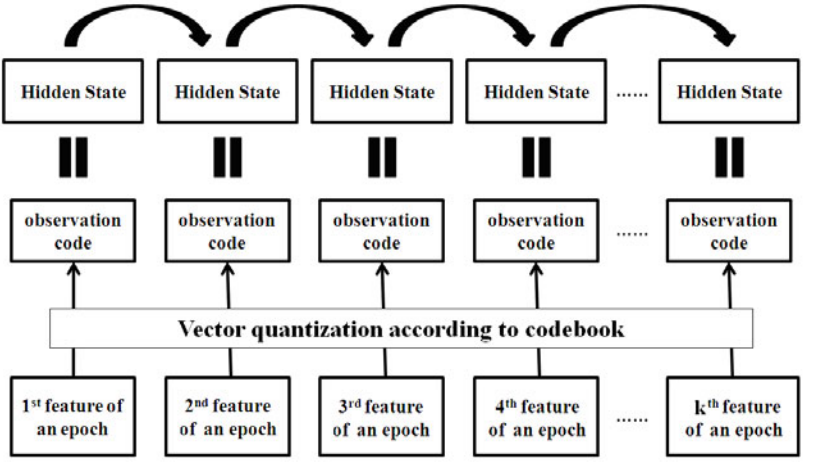


Fig. 2. Relation between the features of speech, the observation and the hidden states of DHMM

To train the DHMM model parameters $\lambda = (A, B, \pi)$ based on existing data, some notations are defined for convenience as follows:

E_{ij} : the event of the transition from state x_i to state x_j

$E_{i\bullet}$: the event of the transition from state x_i to other states

$E_{\bullet j}$: the event of the transition from other states to state x_j

E_{hi} : the event of state x_i appears at initial state

$n(E_{ij})$: the number of the transition from state x_i to state x_j

$n(E_{i\bullet})$: the number of the transition from state x_i to other states

$n(E_{\bullet j})$: the number of the transition from other states to state x_j

$n(E_{\bullet j}, o = v_k)$: the number of enter to state x_j and observation code is v_k

$n(E_{hi})$: the number of the event of state x_i appears at initial state

In the process of training the matrix $A, B,$ and π of DHMM, the hidden states for each observation are estimated first through the initial $A, B,$ and π by using Viterbi

Algorithm. Then the values $n(E_{ij})$, $n(E_{i\bullet})$, $n(E_{\bullet j})$, $n(E_{\bullet j}, o = v_k)$, $n(E_{hi})$ are computed for the whole training data. Subsequently, the elements in matrices A, B, and π are updated as follows.

$$\overline{a_{ij}} = \frac{n(E_{ij})}{n(E_{i\bullet})}, \tag{1}$$

$$\overline{b_j}(k) = \frac{n(E_{\bullet j}, o = v_k)}{n(E_{\bullet j})}, \tag{2}$$

$$\overline{\pi_i} = \frac{n(E_{hi})}{n_{TD}}, \tag{3}$$

where n_{TD} is the number of training data. Using these updated A, B, and π, we run Viterbi Algorithm again. The above steps are repeated until the matrices A, B, and π converge. The training process for a DHMM is then completed. The procedure for recognizing a speech is then depicted in Fig. 3. In the training phase, the DHMM models corresponding to each speech are first trained by using the training speech features through a trained codebook. In the test phase, the feature for a testing speech will be derived first. Through the trained codebook, this feature is then quantized and becomes an observation of the DHMM. For each observation, the probabilities for all DHMM models are calculated. The speech corresponding to the DHMM with largest probability is then the recognized speech.

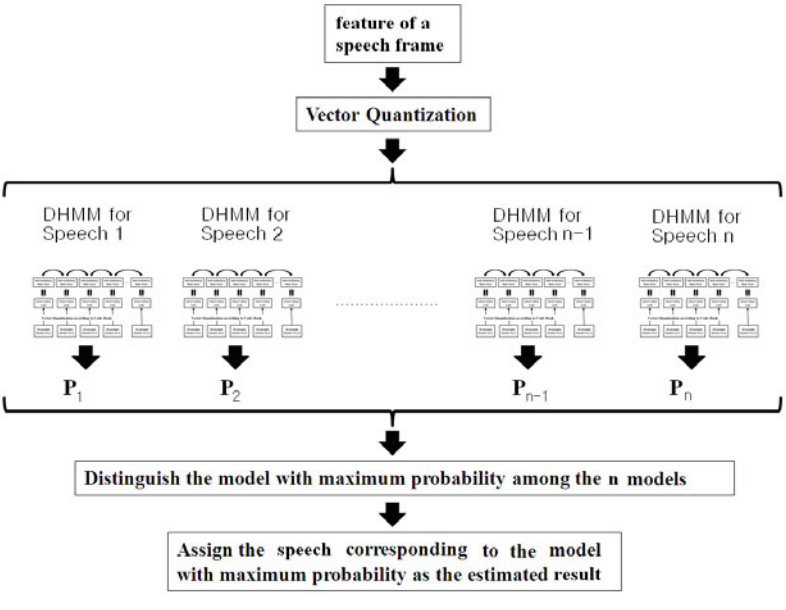


Fig. 3. Procedure for the speech recognition via DHMM

During the speech recognition process, the probability of the observations according to the model $\lambda = (A, B, \pi)$ is calculated by the following equation (4) [13]:

$$\begin{aligned} P(O \mid \lambda) &= \sum_{Q} P(O \mid Q, \lambda) P(Q \mid \lambda) \\ &= \sum_{q_1 \cdots q_T} \pi_{q_1} b_{q_1}(o_1) \cdot a_{q_1 q_2} b_{q_2}(o_2) \cdots a_{q_{T-1} q_T} b_{q_T}(o_T). \end{aligned} \tag{4}$$

This equation enables us to evaluate the probability of the observations O based on the DHMM model $\lambda = (A, B, \pi)$. However, the time taken to evaluate $P(O \mid \lambda)$ directly would be a exponential function of the observation number T. For this reason, the Forward Algorithm [13] is applied to reduce the computation time and is described as follows.

The Forward Algorithm

The Forward Algorithm can be described with three steps, initialization, recursion and termination. The details are listed below based on the above parameters defined for the DHMM.

Initialization: The initial intermediate probabilities $\alpha_1(i)$ for the first observation o_1 are calculated at the beginning as follows.

$$\alpha_1(i) \equiv \pi_i b_i(o_1), \quad 1 \le i \le N. \tag{5}$$

Recursion: For each observation o_t, t = 2, ..., T, the partial probabilities $\alpha_t(j)$ is calculated for each state:

$$\alpha_{t+1}(j) = [\sum_{i=1}^{N} \alpha_t(i) a_{ij}] b_j(o_{t+1}), \quad 1 \le t \le T-1, \quad 1 \le j \le N, \tag{6}$$

in which j is the index number for hidden states. In this step, we calculate the product of the observation probability o_{t+1} and the sum over all possible routes to that state from the states in previous observation o_t. And then, the recursion is performed by using these values from the previous time step.

Termination: Finally, we sum all the partial probabilities at the final time step T to obtain the final result as follows.

$$P(O \mid \lambda) = \sum_{i=1}^{N} \alpha_T(i) \tag{7}$$

The Forward Algorithm reduces the complexity of calculations from $2TN^T$ to N^2T [13].

3 Monitor Process

Figure 4 illustrates the monitor process on Android device. First, a Bluetooth Adapter is used to connect the Android Device with the ECG Signal Monitor. Then, we send the ECG data (raw data, Temperature, Immediate Heartbeat Rate, High Frequency, Low Frequency) to the Bluetooth Handler, which receive and process the data.

Figure 5 (a), 5 (b) and 5 (c) illustrates the processing in the Bluetooth Handler. Figure 5 (d) illustrates the operation of alert dialog window. If user doesn't press the confirm button in 30 sec., a SMS message will be send to the contactor , when the alert dialog show on screen.

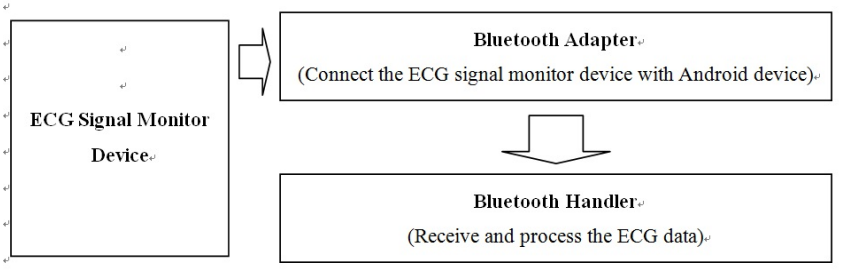


Fig. 4. The Monitor process on Android Device

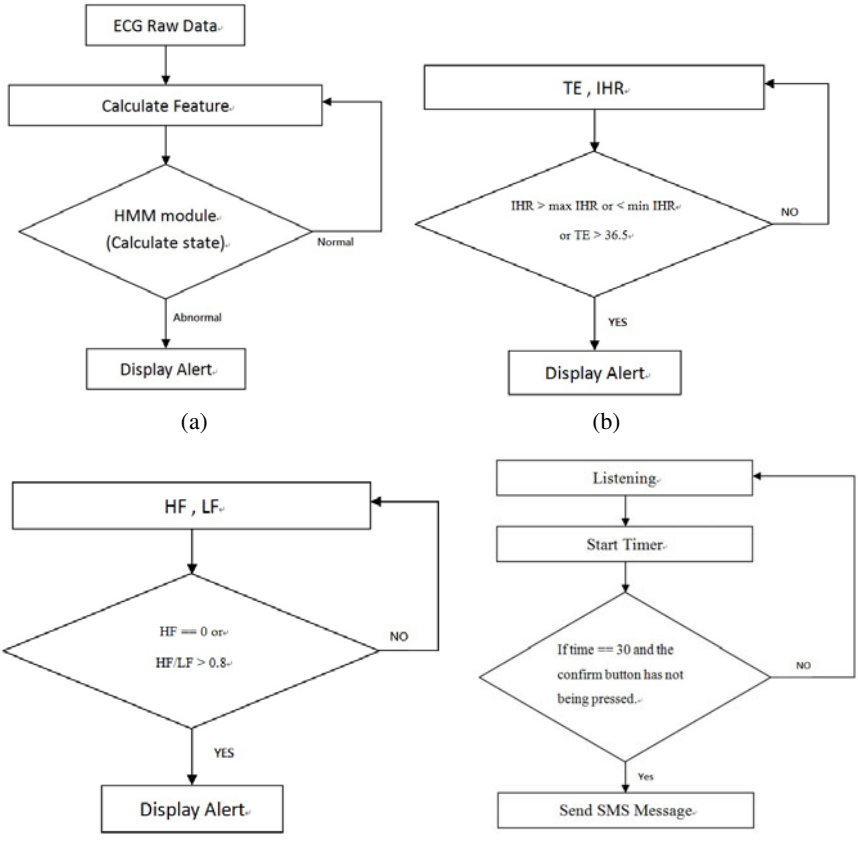


Fig. 5. Processing in the Bluetooth Handler

4 Implementing HMMs

This section reveals the experimental results for the implemented ECG recognition system. We focus on the recognition of ECG by using HMM. We first calculate the features of ECG based on the method proposed in [12]. The features proposed in [12] are shown here in Table 1 and Fig. 6 for convenience. The HMM is first trained according to the algorithm in Section 2 through all ECG features. Then, the trained HMM is used to recognize the feature to distinguish the state of the feature. The data and parameters setting for this experiment are listed as follows.

Data source: MIT-BIH Arrhythmia Database
Signal type: MLII
Data: NORM- 100,103,113,114,122,123,205,234
LBBB- 109,111, 214
RBBB- 118,124, 212, 231
Data length: 50 heart beats for each record

The data for training and testing are listed in Table 2. Moreover, the parameters used in HMM are illustrated as follows.

Hidden state: normal, abnormal
Codebook size: 50×9
Matrix A size: 3×3
Matrix B size: 50×3
Matrix π size: 3×1

Table 3 shows the simulation results in this experiment. The recognition rate for normal heart beats is 100% while for abnormal heart beats is 50%. Hence, the average recognition rate for both normal and abnormal heart beats is about 75%.

Table 1. ECG features in this experiment [12]

Feature's serial No.	Feature's symbol	Feature description.	Units.
1	H-QR	The amplitude between Q and R in a QRS complex.	mV
2	H-RS	The amplitude between R and S in a QRS complex.	mV
3	QRS-dur	The time duration between Q and S in a QRS complex.	ms
4	QTP-int	The time duration between Q and T' in a QRS complex.	ms
5	Ratio-RR	The ratio of *RRs* and *RRa*, *RRs* is the length of a single RR-interval and *RRa* is the average length of all RR-intervals.	-
6	Slope-QR	The slope between Q and R in a QRS complex	mV/ms
7	Slope-RS	The slope between R and S in a QRS complex.	mV/ms
8	Area-QRS	The area of QRS complex.	mV*ms
9	Area-R'ST'	The area of R', S, and T' in a QRS complex. The point R' is the previous point which has the same amplitude as the point T'.	mV*ms

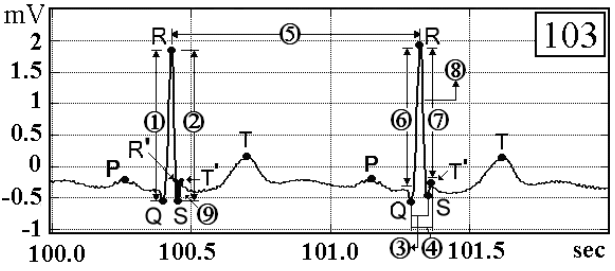


Fig. 6. Waveform of ECG (no. 103) for calculating features [12]

Table 2. Training and testing data

Data	Signal type	Data number
Train data	normal	103,113,114,122,205,234
(total 11 record)	abnormal	109,214,124,212,231
Test data	normal	100,123
(total 4 record)	abnormal	111,118

Table 3. Simulation results

Data type	Data number	**The rate of identity**
normal	**100,123**	**100%**
abnormal	**111,118**	**50%**

5 Conclusions

The system implemented in this paper allows when emergent accidence happened, that ambulance can use the global positioning system in a cellular phone to receive timely rescue. This effectively reduces the abuse of ambulance resources. Ambulance health care system can be best utilized. Users can also attach more different functions in this system, such as diabetes blood glucose monitoring function. Moreover, blood pressure in hypertensive patients can be detected and monitored remotely. More the number of people use this system is, more perfect precision the system is. The use of the system can reduce the communication time delay of for dealing an accident. In the future we can cooperate with hospital to make the system more perfect.

References

1. Rangayyan, R.M.: Biomedical Signal Analysis: A Case-Study Approach. Wiley, Inter-Science, New York (2001)
2. Christov, I., Gómez-Herrero, G., Krasteva, V., Jekova, I., Gotchev, A., Egiazarian, K.: Comparative study of morphological and time-frequency ECG descriptors for heartbeat classification. Med. Eng. Phys. 28, 876–887 (2006)

3. Chazal, P., O'Dwyer, M., Reilly, R.B.: Automatic classification of heart-beats using ECG morphology and heartbeat interval features. IEEE Trans. on Biomed. Eng. 51, 1196–1206 (2004)
4. Throne, R.D., Jenkins, J.M., Dicarlo, L.A.: A comparison of four new time domain techniques for discriminating monomorphic ventricular tachycardia from sinus rhythm using ventricular waveform morphology. IEEE Trans. on Biomed. Eng. 38, 561–570 (1991)
5. Pannizzo, F., Furman, S.: Frequency spectra of ventricular tachycardia and sinus rhythm in human intracardiac electrograms: Application to tachycardia for cardiac pacemakers. IEEE Trans. on Biomed. Eng. 35, 421–425 (1998)
6. Afonso, V.X., Tomkins, W.J., Nguyen, T.Q., Luo, S.: ECG beat detection using filter banks. IEEE Trans. on Biomed. Eng. 46, 192–202 (1999)
7. Benitez, D., Gaydecki, P.A., Zaidi, A., Fitzpatrick, A.P.: The use of the Hilbert transform in ECG signal analysis. Comput. Biol. Med. 31, 399–406 (2001)
8. Koski, A.: Modelling ECG signals with Hidden Markov Models. Artif. Intell. Med. 8, 453–471 (1996)
9. Shahram, M., Nayebi, K.: ECG beat classification based on a Cross-Distance analysis. In: International Symposium on Signal Processing and its Applications, ISSPA 2001, Malaysia, pp. 234–237 (2001)
10. Li, C.W., Zheng, C.X., Tai, C.F.: Detection of ECG characteristic points using wavelet transform. IEEE Trans. Biomed. Eng. 42, 21–28 (1995)
11. Laguna, P., Jane, R., Olmos, S., Thakor, N.V., Rix, H., Caminal, P.: Adaptive estimation of QRS complex wave features of ECG signal by the Hermite model. Med. Biol. Eng. Comput. 34, 58–68 (1996)
12. Yeh, Y.C., Wang, W.J., Chiou, C.W.: Heartbeat Case Determination Using Fuzzy Logic Method on ECG Signals. International Journal 260 of Fuzzy Systems 11(4), 350–361 (2009)
13. Blunsom, P.: Hidden Markov Model. The University of Melbourne, Department of Computer Science and Software Engineering (August 19, 2004)

Reliability Sequential Sampling Test Based on Exponential Lifetime Distributions under Fuzzy Environment

Tien-Tsai Huang[1], Chien-Ming Huang[1], and Kevin Kuan-Shun Chiu[2]

No.300, Sec.1, Wanshou Rd., Gueishan Shiang, Taoyuan County 33306, Taiwan
[1] Department of Industrial Management,
Lunghwa University of Science and Technology
[2] Graduate School of Business and Management,
Lunghwa University of Science and Technology
normanbb@mail.lhu.edu.tw

Abstract. In the areas of reliability analysis, the applications of life testing play a great important role. Under the deterministic circumstances, production reliability acceptance test (PRAT) for reliability sequential-sampling test (RSST) possesses high reliability and economic values in various fields. However, the past literatures showed that some of these uncertainties during the sampling inspection originate from trend, season, cyclic, or random variations; others come from improper operations. Hence, under these factors effect, the RSST may generate a vague value. In order to deal with uncertainty happened; this paper applies the triangular fuzzy number (TFN) are used to express the fuzzy phenomenon of the reliability sampling inspection's parameters. Also, both the centroid and the signed distance approaches are used for defuzzification. To compare these estimates of the reliability sequential-sampling plan in the fuzzy sense, the better defuzzification method is generated.

Keywords: Production Reliability Acceptance Test, Reliability Sampling Plan, Triangular Fuzzy Number, Centroid, Signed Distance.

1 Introduction

Statistical methods and Probability theories are useful in quality control (QC) and reliability analysis (RA). Specifically, it focus on three major areas: statistical process control (SPC), design of experiments (DOE), and acceptance sampling (AS). In addition to these techniques, other statistical tools such as 6-Sigma, Minitab or AMOS they are useful in analyzing quality problems and can be use to improve the performance of production process. This paper focuses on acceptance sampling because the popularity in the fields of manufacturing and servicing businesses. Acceptance sampling was developed in 1941 by H. F. Dodge and H. G. Romig of the Bell telephone laboratories [13]. This technique dating back to long before statistical methodology was developed for quality improvement is closely connected with

J.-S. Pan, S.-M. Chen, N.T. Nguyen (Eds.): ACIIDS 2012, Part III, LNAI 7198, pp. 346–355, 2012.

inspection and testing of product, which is one of the earliest aspects of quality control. The spot for inspection can be set at any time or place during a process. Reliability test, According to the judgment approaches of reliability test results, there are three categories: reliability growth test (RGT), reliability qualification test (RQT) and PRAT [4]. This paper forces on the PRAT.

Observations described or constructed by parametric models of traditional PRAT are precise numbers. For instances, the number of failures *r*, test time *t*, acceptance reliability level (ARL), limiting reliability level (LRL), producer's risk α, and consumer's risk β and so on. We can only get an estimate of the ARL= θ_0, LRL= θ_1, α and β because it is difficult to keep a process in control with a fixed process level for a long period of time. Hence, the α and β should be decided by the designer and the user. Due to human imprecise linguistic expressions, we find it more suitable to express these parameters in fuzzy set concepts. Since the first publication in the fuzzy set theory by Zadeh in 1965 [14], it is over forty years. The fuzzy set theory has been widely applied to many different managerial aspects; such as inventory, decision-making, and other fields. Consequently, the TFN is applied in the *Pa*, α and β. Finally, we compared centroid defuzzification process to that of signed distance when applying the fuzzy set theory to an acceptance test procedure. The successful presentation of the fuzzy PRAT for RSST will be used as an analysis tool in the control of a manufacturing process. Besides, we are liable to judge whether this a better tool when compared with a conventional precise model.

This paper is organized into five sections, section 2 outlines the methods employed. Section 3 describes the research data used in this study. Besides, these data were used to construct the mentioned fuzzy model. In Section 4, we present numerical examples for illustration. As can be noticed, empirical results are recorded, traced, analyzed, reported, and discussed. Section 5 concludes this paper with accomplishments and contributions.

2 Definition

Some basic definitions which consist of the fuzzy set theory and the RSST are introduced in the following sections.

Definition 2.1. The fuzzy set of A represents as $\tilde{A}$. The membership function of $\tilde{A}$ is denoted by $\mu_{\tilde{A}}$, that is $\tilde{A} = \left\{[x, \mu_{\tilde{A}}(x)] \middle| x \in R\right\}$, $0 \le \mu_{\tilde{A}}(x) \le 1$, where x and R represent the elements and universe respectively.

Definition 2.2. According to the Pu and Liu [9,10], let $\tilde{h}$ be a fuzzy set on $R = (-\infty, \infty)$ and $h \in R$, then, $\tilde{h}$ is called a fuzzy point at *h*, if its membership function is given by

$$\mu_{\tilde{h}}(x) = \begin{cases} 1, & x = h, \\ 0, & x \neq h. \end{cases}$$

Let $\tilde{A}$ be a fuzzy set on R, from Zimmermann [9,15], if $\tilde{A} = (a,b,c)$, $a < b < c$, then the membership function of $\mu_{\tilde{A}}(x)$ is defined as

$$\mu_{\tilde{A}}(x)=\begin{cases}(x-a)/(b-a), a\le x\le b,\\ (c-x)/(c-b), b\le x\le c,\\ 0, \quad otherwise.\end{cases} \tag{1}$$

Geometrically, Eq. (1) defines the triangular fuzzy number (TFN).

Definition 2.3. Suppose a fuzzy set $\tilde{A}$ is given and for any $\alpha\in[0,1]$. a crisp set $A_\alpha=\{\mu_{\tilde{A}}(x)\ge\alpha | x\in R\}$ is defined, which is called α-cut of $\tilde{A}$. The α-cuts of A_α is a closed interval, $A_\alpha=[A_L(\alpha),A_R(\alpha)]$, where $A_L(\alpha)$ and $A_R(\alpha)$ are the left point and right point of interval A_α respectively.

Definition 2.4. From Klir et al. and Kaufmann and Gupta [1,2], the following operations are obtained.

$$A_\alpha(-)B_a=\bigcup_{0\le\alpha\le1}[(a-r)+((b-a)-(r-q))\alpha,(c-p)-((c-b)-(q-p))\alpha;\alpha]. \tag{2}$$

$$A_\alpha(:)B_a=\bigcup_{0\le\alpha\le1}[(a+(b-a)\alpha)/(r-(r-q)\alpha),(c-(b-c)\alpha)/(p+(q-p)\alpha);\alpha]. \tag{3}$$

$$kA_\alpha=\bigcup_{0\le\alpha\le1}[ka+k(b-a)\alpha,kc-k(b-c)\alpha;\alpha]. \tag{4}$$

$$A_\alpha^{-1}=\bigcup_{0\le\alpha\le1}[c^{-1}-(c^{-1}-b^{-1})\alpha,a^{-1}+(b^{-1}-a^{-1})\alpha;\alpha]. \tag{5}$$

$$\ln A_\alpha=\bigcup_{0\le\alpha\le1}[\ln a+\alpha\ln(b/a),\ln c-\alpha\ln(c/b);\alpha]. \tag{6}$$

Definition 2.5. Let $\tilde{A}=(a,b,c)$ be a TFN and $C(\tilde{A})$ be denoted as the value of defuzzification of $\tilde{A}$ by centroid [3], which is given by

$$C(\tilde{A})=\left[\int_{-\infty}^{\infty}x\mu_{\tilde{A}}(x)dx\right]\Big/\left[\int_{-\infty}^{\infty}\mu_{\tilde{A}}(x)dx\right]=\frac{1}{3}(a+b+c). \tag{7}$$

Definition 2.6. Let $\tilde{A}=(a,b,c)$ be a TFN and $d(\tilde{A},\tilde{0})$ be denoted as the value of defuzzification of $\tilde{A}$ by signed distance [12], which is given by

$$d(\tilde{A},\tilde{0})=\frac{1}{2}\int_0^1[A_L(\alpha)+A_R(\alpha)]d\alpha=\frac{1}{4}(2b+a+c). \tag{8}$$

Definition 2.7. The life test can be divided two situations; the replacement (mean time between failure; MTBF) and non-replacement (mean time to failure; MTTF) [7]. Suppose we have *n* samples for conducting life test of non-replacement, then the total

test time that the failure occurred until r^{th} time is $T=\sum_{i=1}^{r}t_i+(n-r)t_r$. On the contrary, if we conduct life test of replacement, then the total test time that the failure occurred until r^{th} time is $T=\sum_{i=1}^{r}t_i=nt_r$.

Definition 2.8. The operating characteristic (OC) curve of PRAT for MTBF(θ) in different situation to plot the Pa curves is illustrated in Fig. 1 [4,7]:

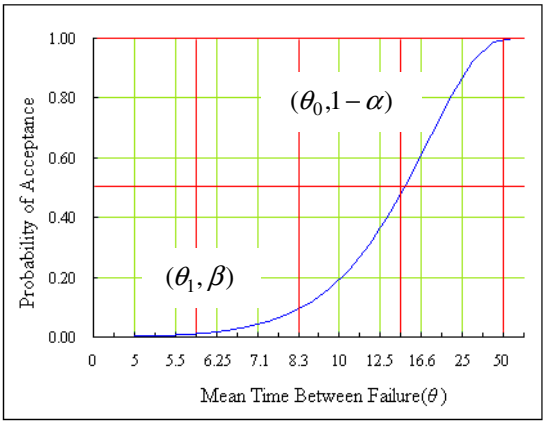


Fig. 1. The OC curve of MTBF(θ)

Definition 2.9. Suppose that we eish to construct a sampling plan such that the $Pa(\theta_0)$ is $1-\alpha$ for lots with ARL $=\theta_0$, and the $Pa(\theta_1)$ is β for lots with LRL $=\theta_1$. Assuming poisson sampling is appropriate, the number of failures r and acceptance number of failures $k=r_0-1$ was obtained by writing out the two point on the OC curve [6,7]. The following definitions of two equations are given by

$$P_0(r\le r_0-1|\theta=\theta_0)=Pa(\theta_0)=\sum_{r=0}^{k=r_0-1}\frac{1}{r!}\left(\frac{T}{\theta_0}\right)^r e^{-\frac{T}{\theta_0}}=1-\alpha. \tag{9}$$

$$P_1(r\le r_0-1|\theta=\theta_1)=Pa(\theta_1)=\sum_{r=0}^{k=r_0-1}\frac{1}{r!}\left(\frac{T}{\theta_1}\right)^r e^{-\frac{T}{\theta_1}}=\beta. \tag{10}$$

Definition 2.10. RSST is based on the probability ratio sequential test (PRST), developed by Wald [11], from Eqs. (9) and (10), the PRST is defined by

$$\text{PRST}=\frac{P_1(r<k|\theta=\theta_1)}{P_0(r<k|\theta=\theta_0)}=\left(\frac{\theta_0}{\theta_1}\right)^r e^{-T\left(\frac{1}{\theta_1}-\frac{1}{\theta_0}\right)}.$$

The operation of a RSST is illustrated in Fig. 2. The cumulative observed number of failures is plotted on the chart. For each point, the abscissa is the total test time selected up to that time, and the ordinate is the total number of observed failures.

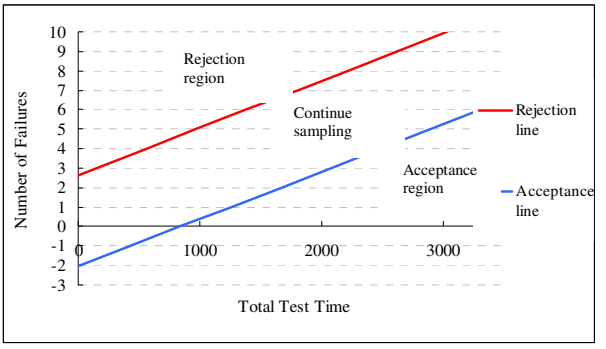


Fig. 2. The RSST

Definition 2.11. Determine the discrimination ratio $d = \theta_0 / \theta_1$. Let the sequential tests are all truncated tests because of the practical requirements of real-world test programs. The method for truncating a sequential test was developed by Epstein and Sobel [8]. The appropriate value of *r* is the smallest integer that is given by

$$\frac{\chi^2_{1-\alpha,2r}}{\chi^2_{\beta,2r}} \geq \frac{\theta_1}{\theta_0}. \tag{11}$$

where $\chi^2_{1-\alpha,2r}$ and $\chi^2_{\beta,2r}$ are the chi-square variables with $2r$ degrees of freedom (*df*) These two values are found by simultaneously searching the $1-\alpha$ and β probabilities of the chi-square tables until the ratio of the variables is equal to or greater than θ_1 / θ_0. While this point is found, the *df* is $2r$. The value of *r* is always rounded to the next highest integer, and the value is r_0. Fig. 3 shows the truncated number of failures for RSST.

From this truncated number of failures is r_0. the maximum time T_0 can be found. Fig. 4 shows the truncated test time for RSST.

$$T_0 = \frac{\theta_0 \chi^2_{1-\alpha,2r_0}}{2} = \frac{\theta_1 \chi^2_{\beta,2r_0}}{2}. \tag{12}$$

3 RSST with Fuzzy Parameter

In traditional method, the PRAT is the test that compares the test results and the setting value via the statistical technique to determine whether the product reliability matches the requirement or not. In the mean time, the test results are also used for rating the product capacity. In recent years, the fuzzy sets theory has been applied on reliability theory. In the following, the processes of reliability sequential-sampling test with fuzzy parameter are depicted by three steps.

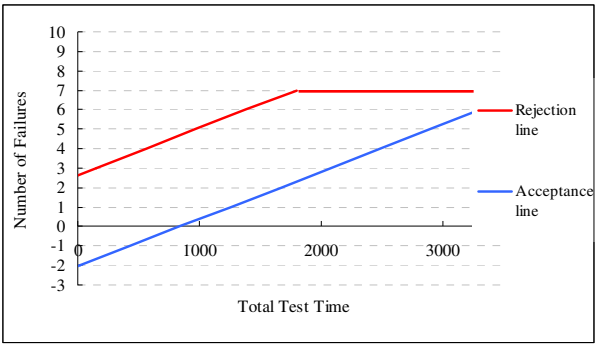


Fig. 3. The truncated number of failures for RSST

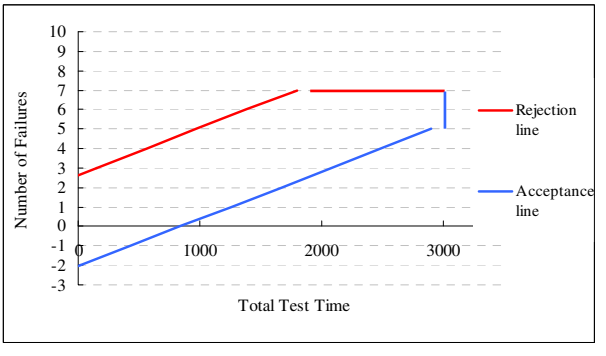


Fig. 4. The truncated test time for RSST

Step1: First we consider they are both precise numbers. Based on above assumptions, the fuzzy set theory is applied in RSST's operate parameters. The membership function of producer's risk $\mu_{\tilde{\alpha}}(u)$, and consumer's risk $\mu_{\tilde{\beta}}(v)$. which is called fuzzy designer's risk and fuzzy user's risk. Let $\tilde{\alpha} = (\alpha - \Delta_1, \alpha, \alpha + \Delta_2)$, $0 < \Delta_1 < \alpha$, $0 < \Delta_2$ and $\tilde{\beta} = (\beta - \Delta_3, \beta, \beta + \Delta_4)$, $0 < \Delta_3 < \beta$, $0 < \Delta_4$ be TFNs and following Eq. (1), then the $\mu_{\tilde{\alpha}}(u)$ and $\mu_{\tilde{\beta}}(v)$ are expressed as follows

$$\mu_{\tilde{\alpha}}(u) = \begin{cases} \dfrac{u - (\alpha - \Delta_1)}{-\Delta_1}, \alpha - \Delta_1 \le u \le \alpha, \\ \dfrac{(\alpha + \Delta_2) - u}{\Delta_2}, \alpha \le u \le \alpha + \Delta_2, \\ 0, \qquad otherwise. \end{cases}$$

and

$$\mu_{\tilde{\beta}}(v) = \begin{cases} \dfrac{v - (\beta - \Delta_3)}{-\Delta_3}, \beta - \Delta_3 \le v \le \beta, \\ \dfrac{(\beta + \theta_4) - v}{\Delta_4}, \beta \le v \le \beta + \Delta_4, \\ 0, \qquad otherwise. \end{cases}$$

Than the membership function of ARL $\mu_{\tilde{\theta}_0}(\vartheta_0)$, and LRL $\mu_{\tilde{\theta}_1}(\vartheta_1)$. which is called fuzzy acceptance reliability level (FARL) and fuzzy limiting reliability level (FLRL). Let $\tilde{\theta}_0 = (\theta_0 - \nabla_1, \theta_0, \theta_0 + \nabla_2)$, $0 < \nabla_1 < \alpha$, $0 < \nabla_2$ and $\tilde{\theta}_1 = (\theta_1 - \nabla_3, \theta_1, \theta_1 + \nabla_4)$, $0 < \nabla_3 < \beta$, $0 < \nabla_4$ be TFNs and following Eq. (1), then the $\mu_{\tilde{\theta}_0}(\vartheta_0)$ and $\mu_{\tilde{\theta}_1}(\vartheta_1)$ are expressed as follows

$$\mu_{\tilde{\theta}_0}(\vartheta_0) = \begin{cases} \dfrac{\vartheta_0 - (\theta_0 - \nabla_1)}{-\nabla_1}, \theta_0 - \nabla_1 \le \vartheta_0 \le \theta_0, \\ \dfrac{(\theta_0 + \nabla_2) - \delta}{\nabla_2}, \theta_0 \le \vartheta_0 \le \theta_0 + \nabla_2, \\ 0, \qquad otherwise. \end{cases}$$

and

$$\mu_{\tilde{\theta}_1}(\vartheta_1) = \begin{cases} \dfrac{\vartheta_1 - (\theta_1 - \nabla_3)}{-\nabla_3}, \theta_1 - \nabla_3 \le \vartheta_1 \le \theta_1, \\ \dfrac{(\theta_1 + \theta_4) - \vartheta_1}{\nabla_4}, \theta_1 \le \vartheta_1 \le \theta_1 + \nabla_4, \\ 0, \qquad otherwise. \end{cases}$$

Step 2: In order to plot six straight lines, i.e., fuzzy acceptance line (FAL) and fuzzy rejection line (FRL), two TFNs' operations of the fuzzy RSST's are conducted respectively. Let $\tilde{T}r = (Tr_1, Tr_2, Tr_3)$ and $\tilde{T}a = (Ta_1, Ta_2, Ta_3)$ be TFNs respectively, where $Tr_1 < Tr_2 < Tr_3$, $Ta_1 < Ta_2 < Ta_3$. The computed results are shown in Fig. 5.

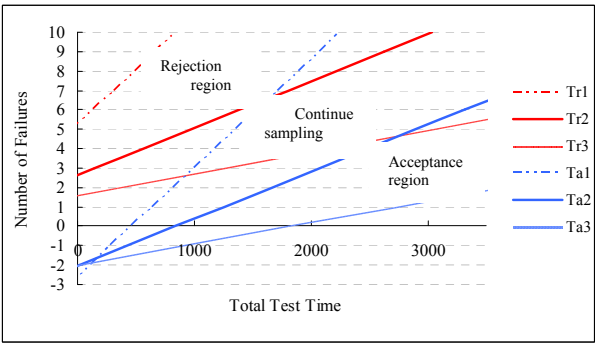


Fig. 5. The fuzzy RSST

Suppose one plotted point is sampling, then three situations are occurred. When a point falls on the area which is above the upper line Tr_1 (locates at the rejection region), then the lot is rejected. If a point falls on the area which is below the lower line Tr_3 (locates at the continuing sampling region), then the RSST continues the plotted point locates at the other two regions. Hence, the FRL is denoted by

$$\tilde{T}r \ge \left[\ln\left(\frac{1-\tilde{\beta}}{\tilde{\alpha}}\right)(-) r \ln\left(\frac{\tilde{\theta}_0}{\tilde{\theta}_1}\right) \right](:)\left(\frac{1}{\tilde{\theta}_1}(-)\frac{1}{\tilde{\theta}_0}\right) = (Tr_1, Tr_2, Tr_3). \tag{13}$$

where

$$(Tr_1,Tr_2,Tr_3)=\left[\frac{\ln\left(\frac{1-\beta+\Delta_4}{\alpha+\Delta_2}\right)-r\ln\left(\frac{\theta_0+\nabla_2}{\theta_1-\nabla_3}\right)}{\left(\frac{1}{\theta_1-\nabla_3}-\frac{1}{\theta_0+\nabla_2}\right)},\frac{\ln\left(\frac{1-\beta}{\alpha}\right)-r\ln\left(\frac{\theta_0}{\theta_1}\right)}{\left(\frac{1}{\theta_1}-\frac{1}{\theta_0}\right)},\frac{\ln\left(\frac{1-\beta-\Delta_3}{\alpha-\Delta_1}\right)-r\ln\left(\frac{\theta_0-\nabla_1}{\theta_1+\nabla_4}\right)}{\left(\frac{1}{\theta_1+\nabla_4}-\frac{1}{\theta_0-\nabla_1}\right)}\right].$$

Similarly, when a point falls on or below the lower line Ta_3 (locates at the acceptance region), the lot is accepted. If a point falls on or above the upper line Ta_1 (locates at the continuing sampling region). Hence, the FAL is given by

$$\tilde{T}a\le\left[\ln\left(\frac{\tilde{\beta}}{1-\tilde{\alpha}}\right)(-)r\ln\left(\frac{\tilde{\theta}_0}{\tilde{\theta}_1}\right)\right](:)\left(\frac{1}{\tilde{\theta}_1}(-)\frac{1}{\tilde{\theta}_0}\right)=(Ta_1,Ta_2,Ta_3), \tag{14}$$

where

$$(Ta_1,Ta_2,Ta_3)=\left[\frac{\ln\left(\frac{\beta-\Delta_3}{1-\alpha-\Delta_1}\right)-r\ln\left(\frac{\theta_0+\nabla_2}{\theta_1-\nabla_3}\right)}{\left(\frac{1}{\theta_1-\nabla_3}-\frac{1}{\theta_0+\nabla_2}\right)},\frac{\ln\left(\frac{\beta}{1-\alpha}\right)-r\ln\left(\frac{\theta_0}{\theta_1}\right)}{\left(\frac{1}{\theta_1}-\frac{1}{\theta_0}\right)},\frac{\ln\left(\frac{\beta+\Delta_4}{1-\alpha+\Delta_2}\right)-r\ln\left(\frac{\theta_0-\nabla_1}{\theta_1+\nabla_4}\right)}{\left(\frac{1}{\theta_1+\nabla_4}-\frac{1}{\theta_0-\nabla_1}\right)}\right].$$

However, if the plotted point locates continuously at the interval $[Tr_1,Tr_3]$ and $[Ta_1,Ta_3]$ twice, then the lot is rejected or accepted.

Step 3: Using centroid for defuzzification of $\tilde{\theta}_0$ and $\tilde{\theta}_1$, then we have $C(\tilde{\theta}_0)=(3\theta_0-\nabla_1+\nabla_2)/3$ and $C(\tilde{\theta}_1)=(3\theta_1-\nabla_3+\nabla_4)/3$, which decide the value of defuzzification by centroid for discrimination ratio $d_c=C(\tilde{\theta}_0)/C(\tilde{\theta}_1)$. Similarly, using signed distance to defuzzy $\tilde{\theta}_0$ and $\tilde{\theta}_1$ respectively, then we have $d(\tilde{\theta}_0,\tilde{0})=(4\theta_0-\nabla_1+\nabla_2)/4$ and $d(\tilde{\theta}_1,\tilde{0})=(4\theta_1-\nabla_3+\nabla_4)/4$, which decide the value of defuzzification by signed distance for discrimination ratio $d_s=d(\tilde{\theta}_0,\tilde{0})/d(\tilde{\theta}_1,\tilde{0})$.

The value of defuzzification by centroid and signed distance for appropriate value of *r* is the smallest integer which can be given by

$$\frac{\chi^2_{1-\alpha,2r}}{\chi^2_{\beta,2r}}\ge\frac{C(\tilde{\theta}_1)}{C(\tilde{\theta}_0)}\approx\frac{d(\tilde{\theta}_1,\tilde{0})}{d(\tilde{\theta}_0,\tilde{0})} \tag{15}$$

Due to the value of defuzzification of truncated number of failures is r_0, from the Eqs. (15)-(16) for defuzzification of truncated test time by centroid and signed distance, the maximum time T_c and T_s is given by

$$T_c=\frac{1}{2}C(\tilde{\theta}_0)\chi^2_{1-\alpha,2r_0}\approx\frac{1}{2}C(\tilde{\theta}_1)\chi^2_{\beta,2r_0}. \tag{16}$$

$$T_s=\frac{1}{2}d(\tilde{\theta}_0,\tilde{0})\chi^2_{1-\alpha,2r_0}\approx\frac{1}{2}d(\tilde{\theta}_1,\tilde{0})\chi^2_{\beta,2r_0}. \tag{17}$$

4 Numerical Examples

To illustrate the operational feasibility of these approaches, a data set were obtained from fuzzy PRAT for RSST, and applied in this section. The purpose of using these data is to compare the results among the different approaches. If we have fuzzy RSST where $\tilde{\theta}_0 = (600,750,900)$, $\tilde{\alpha} = (0.025,0.05,0.075)$, $\tilde{\theta}_1 = (200,250,300)$, and $\tilde{\beta} = (0.05,0.1,0.15)$, therefore, Using Eqs. (13) and (14), we see that the value of FRL and FAL for the RSST are $\tilde{T}r = (-1456.65+178.24r,-1083.89+411.98r,-935.38+902.45r)$ and $\tilde{T}a = (467.78+178.24r,844.23+411.98r,1812.63+902.45r)$ respectively.

Instead of using a graph to determine the lot disposition, the fuzzy RSST can be displayed in a Table 1. The entries in the table are found by substituting values of n into the equations for the FAL, FRL and calculating acceptance and rejection number of failures.

Table 1. Fuzzy RSST

r	Rejection test time	Acceptance test time
0	(*N/A*, *N/A*, *N/A*)	(467.78, 844.23, 1812.63)
1	(*N/A*, *N/A*, *N/A*)	(646.02, 1256.21, 2715.08)
2	(*N/A*, *N/A*, 348.24)	(824.26, 1668.19, 3617.52)
3	(*N/A*, 152.05, 1250.69)	(1002.5, 2080.17, 4519.97)
4	(*N/A*, 564.03, 2153.14)	(1180.73, 2492.15, 5422.42)
5	(*N/A*, 976.01, 3055.58)	(1358.97, 2904.13, 6324.86)
6	(134.05, 1387.99, 3958.03)	(1537.21, 3316.11, 7227.31)
7	(312.29, 1799.97, 4806.48)	(1715.45, 3728.09, 8129.76)
8	(490.52.78, 2211.95, 5762.92)	(1893.69, 4140.07, 9032.2)
9	(668.76, 2623.93, 6665.37)	(2071.92, 4552.05, 9934.65)
10	(846.99, 3035.91, 7567.82)	(2250.16, 4964.03, 10837.1)

Therefore, the defuzzification of discrimination ratio *d* by centroid and signed distance, from the step 3, we have $d_c = 3.166$ and $d_s = 3.125$, where $\chi^2_{0.95,16}$ and $\chi^2_{0.1,16}$ are the chi-square variables with the $df = 16$. When this point is found, the defuzzification of truncated number of failures is 8. Hence, the number of failures for acceptance is 7 and the value of defuzzified of truncated number of failures is r_0. From the Eqs. (16)-(17) for defuzzification of truncated test time, The maximum time $T_c = 2853.05$ and $T_s = 2886.23$ are generated respectively. According to the results, the discrimination ratio *d,* truncated number of failures and test time based on centroid is better than that of signed distance. Table 2 shows the results.

Table 2. Defuzzification of the d, r_0 and T_0

	Centroid	Signed distance
d	3.166*	3.125
r_0	8	8
T_0	2853.05*	2886.23

5 Conclusions

In this paper, the fuzzy set theory with TFNs was applied to acceptance test. Authors not only used membership functions to express the fuzzy phenomenon of measured values but also proposed an uncertain fuzzy number of failures sampled from the PRAT for conducting a RSST. Several research questions have been examined and researched. Major results and findings are described as follows:(1) This paper reveals possible ways of using fuzzy set theory in dealing with acceptance test problems under uncertain operation conditions; (2) The applied fuzzy concepts include the constructing the membership function of fuzzy parameter and specifying the fuzzy PRAT for a RSST in vague situations; (3) When compared with the traditional PRAT, the fuzzy PRAT needs more number of failures. Hence, the fuzzy PRAT has to pay a higher sampling cost and longer test time. However, the contribution of this approach can be used to enhance the stability of a manufacturing process.

References

1. Kaufmann, A., Gupta, M.M., Esposito, B.: Introduction to Fuzzy Arithmetic: Theory and Applications. Van Nostrand Reinhold Company (1991)
2. Klir, G.J., Yuan, B.: Fuzzy Sets and Fuzzy Logic. Prentice Hall PTR, Upper Saddle River (1995)
3. Klir, G.J., Clair, U.S., Yuan, B.: Fuzzy Sets Theory-Foundations and Applications. Prentice Hall PTR, Upper Saddle River (1997)
4. Ko, H.-Y.: Reliability Assurance, Chinese Society for Quality (2008)
5. Lee, Y.-C., Wang, H.-F., Su, M.-C.: Fuzzy Theory and Application. Chuan Hwa Book CO., LTD. (2008)
6. Nelson, W.: Applied Life Data Analysis. John Wiley and Sons, New York (1982)
7. Wang, T.-H.: Reliability Engineering and Management. Chinese Society for Quality (2010)
8. Epstein, B., Sobel, M.: Life Testing. Journal of the American Statistical Association 48(263), 486–502 (1953)
9. Huang, T.-T.: Some Applications of Fuzzy Evaluation System for Service Quality and Product Lot-Sizing. Dept. of Management Sciences. TamKang University (2006)
10. Pu, P.M., Liu, Y.M.: Fuzzy Topology 1, Neighborhood Structure of a Fuzzy point and Moore-Smith Convergence. Journal of Mathematics Analysis and Applications 76, 571–599 (1980)
11. Wald, A.: Sequential Analysis. Wiley, New York (1947)
12. Yao, J.S., Wu, K.M.: Ranking Fuzzy Number Based on Decomposition Principle and Distance. Fuzzy Sets and Systems 116, 275–288 (2000)
13. Yang, H.-P.: Process Capability Indices with Fuzzy Numbers. Dept. of Statistics. National ChengKung University (2006)
14. Zadeh, L.A.: Fuzzy set. Information and Control 8, 338–353 (1965)
15. Zimmermann, H.J.: Fuzzy Set Theory and Its Applications, 3rd edn. Kluwer Academic Publishers, Boston (1996)

Adaptive Performance for VVoIP Implementation in Cloud Computing Environment

Bao Rong Chang[1,*], Hsiu-Fen Tsai[2], Zih-Yao Lin[1],
Chi-Ming Chen[1], and Chien-Feng Huang[1]

[1] Department of Computer Science and Information Engineering
National University of Kaohsiung, Kaohsiung, Taiwan
{brchang,cfhuang15}@nuk.edu.tw,
qazwsxee3118@xuite.net, gcsul0725@hotmail.com
[2] Department of Marketing Management
Shu Te University, Kaohsiung, Taiwan
soenfen@stu.edu.tw

Abstract. In this paper we have implemented a real-time video/voice over IP (VVoIP) applications on a Hadoop cloud computing system and it is denoted CLC-IHU. It really outperforms the previous VVoIP using P2P connection (called SCTP-IHU) due to the easy-to-use and high-performance on video phone call. User does not need to know what is a real IP and web interface achieves interaction by adopt TCP instead of PR-SCTP so taht CLC-IHU scheme reduces computation load and power consumption dramatically at thin clients. We employed adaptive network-based fuzzy inference system (ANFIS) to tune key factors appropriately for adapting handoff and analyzing network traffic at any time. As a result it takes about 1.631 sec for the seamless handoff between base stations under mobile wireless network. In access control for preventing illegal intrusions from the outside of the cloud, the rapid facial recognition and fingerprint identification via cloud computing has been done successfully within 2.2 seconds to identify the subject exactly.

Keywords: Video/Voice over IP (VVoIP), Hadoop Cloud Computing, ANFIS, PR-SCTP, SCTP-IHU, CLC-IHU.

1 Introduction

It is well-known that audio stream transmitted are through Real-Time Transport Protocol (RTP) [1] relying on UDP, which provides no loss recovery (unreliable). Using TCP rather than UDP, it may cause long delay to get obsolete data at receiver side. In the previous work [2], VVoIP is constructed in ARM-based embedded platform rather than the earlier project implemented on x86 PC [3] for the purpose of the use of mobile video phone call operated on P2P connection [4]. In order to tackle three crucial problems: head-of-line blocking, handover interruption, and non-real-time

* Corresponding author.

J.-S. Pan, S.-M. Chen, N.T. Nguyen (Eds.): ACIIDS 2012, Part III, LNAI 7198, pp. 356–365, 2012.

transmission, we have adopted PR-SCTP protocol [5] instead of TCP. However, the computation load for real-time video phone call is so big and has caused ARM-based embedded platform [6] no longer with power-saving. Moreover, VVoIP need to know their respective IP address between two mobile devices before P2P connection.

In order to resolve the problems as mentioned above, this study is to realize a real-time video/voice over IP (VVoIP) applications that has implemented by a Hadoop cloud computing and it is denoted CLC-IHU. Next based on client/server architecture users employ web interface to achieve VVoIP connection with thin clients and video phone call adopts an easy TCP connection instead of the complicated PR-SCTP protocol because of web interface not supported by PR-SCTP. User does not need to know what is a real IP used to ring a video phone call. In addition, cloud computing server has been protected by the access control according to the mechanism of authentication, authorization, and accounting (AAA) [7]. Hadoop cloud computing [8] together with access control by employing the rapid identification of face and fingerprint has been realized in order for preventing illegal intrusions from the outside of the cloud. Here, we use the standard J2ME [9] environment for embedded devices, where JamVM [10] virtual machine is employed to achieve J2ME environment and GNU Classpath [11] is used as the Java Class Libraries. In terms of quality of services (QoS) concerned in the video phone call over IP, we have to tackle some of crucial problems about jitter, loss, latency, and throughput so as to maintain the smooth video/voice streaming over internet. We employed adaptive network-based fuzzy inference system (ANFIS) [12] to tune key factors appropriately for adapting handoff and analyzing network traffic at any time.

2 Video/Voice over IP on Cloud Computing

Cloud computing is an emerging and increasingly popular computing paradigm, which provides the users massive computing, storage, and software resources on demand. How to program efficient distributed parallel applications is complex and difficult. How to dispatch a large scale task executed in cloud computing environments is challenging as well. Currently on the market the most popular cloud computing services are divided into public clouds, private clouds, community/open clouds, and hybrid clouds, where Goggle App Eng [13], Amazon Web Services [14], Microsoft Azure [15] - the public cloud; IBM Blue Cloud [16] - the private cloud; Open Nebula [17], Eucalyptus [18], Yahoo Hadoop [19] and the NCDM Sector/Sphere [20] - open cloud; IBM Blue Cloud [16] - hybrid cloud.

Two famous companies related to IP phone via cloud computing are voice over IP in a Cloud, IIS 2009 [21] and VoIP in cloud computing, Skpye in 2010 [22]. In Taiwan, Chunghwa Telecom provides Hicloud with two services: CaaS and StaaS [23], Innovative DigiTech-Enabled Applications & Services Institute at III [24] gives public cloud computing services, National Center for cloud computing at NARL delivers training courses for cloud computing [25], and Cloud computing Center for Mobile Application at ITRI contributes the efforts in Container Computer, Cloud OS, and Application [26].

In this study, the VVoIP application program does not need to be installed inside mobile devices. It has been set up in the Hadoop cloud computing server [27] [28] so that web interface is applicable to run and browse video phone call between cloud computing server and thin clients as shown in Fig. 1. Therefore, instead of complicate PR-SCTP protocol a simple and easy control transmission protocol TCP is employed in such a client/server structure rather than P2P connection as we did before. Video/voice streaming over IP can be implemented in such a bidirectional connection between thin-client type of Linux or WinCE embedded platforms. Users do not need to get both actual IPs in advance; instead they get into Hadoop cloud computing server to catch VVoIP service directly for initiating video phone call connection and cloud computing will be looking for the other side to accomplish phone call link between two client sites.

(a) symbol (b) architecture

Fig. 1. Hadoop cloud computing server linked to mobile devices over WiFi

3 Realization of Cloud Computing System

3.1 Deploying Hadoop Cloud Computing

Once a Hadoop cloud computing server has been established in server site, we have to test the functionality of cloud computing in Hadoop system as shown in Fig. 2, 3, and 4. In order to setup a programming environment for Python or Java, an Eclipse IDE [29] is applied to develop the application program (AP) at local site. It is noted that please remember to install Java JDK [30] before you setup an Eclipse IDE in local site. If AP has been done and is waiting for dispatching itself to Hadoop cloud computing server, we deploy this AP via the path of LAN or WiFi. Finally we take a look at HBase in Hadoop server to make sure that the cloud computing is ready for the task.

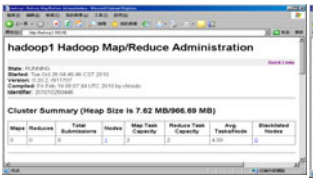


Fig. 2. Testing Hadoop Administration Interface at http://hadoop1:50030

Fig. 3. Testing Task Tracker Status at http://hadoop1:50060

Fig. 4. Testing HDFS Status at http://hadoop1:50070

3.2 Establishing Thin Client and Installing Access Control

In terms of thin client, JamVM is treated as the framework of programming development; however the virtual machine JamVM has no way to perform the drawing even through their core directly, and thus it must call other graphics library to achieve the drawing performance. Here some options we have are available, for example, GTK+DirectFB, GTK+X11, QT/Embedded [31]. As shown in Fig. 5, this study has chosen QT/Embedded framework instead of GTK series, in such a way that achieves

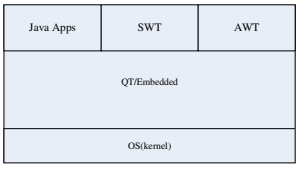


Fig. 5. Testing HDFS Status at http://hadoop1:50070

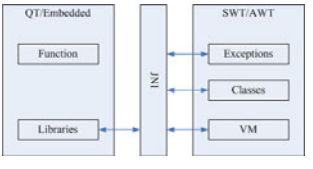


Fig. 6. Communication between SWT/AWT and QT/Embedded

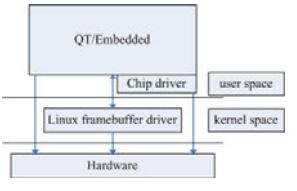
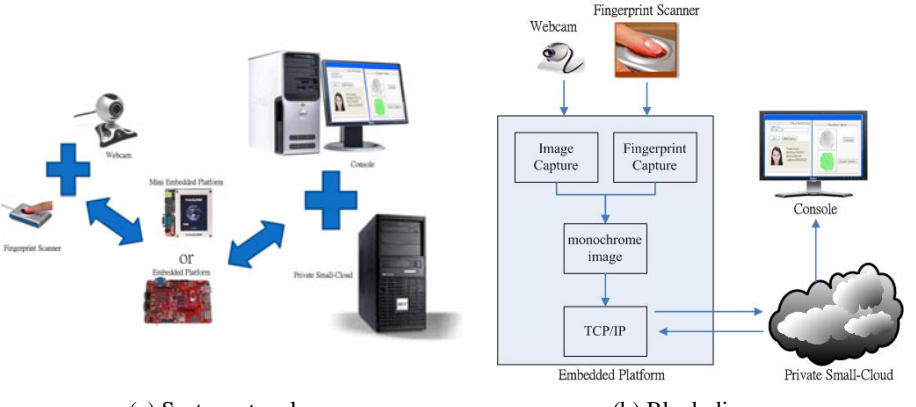


Fig. 7. QT/embedded communicates with the Linux Framebuffer



(a) System at a glance (b) Block diagram

Fig. 8. Architecture of access control

GUI interface functions. In Fig. 6, no matter SWT or AWT in JamVM they apply Java Native Interface (JNI) [32] to communicate C- written graphics library. Afterward QT/Embedded gets through the kernel driver to activate graphic function as shown in Fig. 7. Cloud computing here can perform rapid fingerprint identification [33] and facial recognition [34] in order to fulfill the access control system [35]. We will see whether or not a quick response to client is confirmed. The access control system is shown in Fig. 8.

3.3 Implementing VVoIP at Hadoop with Web Interface

(1) According to a open source IHU (a voice over IP application program), we have established a video/voice over IP application program implemented on Linux system to realize VVoIP over two PCs as shown in Fig. 9.
(2) Activating a database in a cloud computing server to record the necessitated information like account, IP address and so on.
(3) Constructing the web services in order to connect a database in Hadoop cloud computing server for implementing the VVoIP applications, user account and login, as shown in Figs. 10 and 11.
(4) Transplanting the video/voice over IP application program into cloud computing server incorporation with web interface and database.

Fig. 9. VVoIP implementation between two PCs via P2P connection

Fig. 10. Web registration at Hadoop cloud computing server before launching VVoIP application

Fig. 11. Web login interface at Hadoop cloud computing server before launching VVoIP application

3.4 Intelligent Adaptation for VVoIP on Cloud Computing

In terms of quality of services (QoS) concerned in the video phone call over IP, we have to tackle some of crucial problems about jitter, loss, latency, and throughput so as to maintain the smooth video/voice streaming over internet. This study herein introduces a way of intelligent adaptation for key factors to tune video/audio parameters appropriately in Hadoop cloud computing system while an on-line video phone call between clients. As shown in Fig. 12 the diagram draws a picture to show you a

intelligent computation using adaptive network-based fuzzy inference system (ANFIS) [12] to adjust the time for frame delay while video streaming is transmitting and receiving at the same time each other. Besides, the same scheme of ANFIS is applied to audio parameters tuning for the size of video buffer and the least volume in dB for receiver as shown in Fig. 13. We have collected a lot of data by the manner of trial-and-error during the experiments. Once the data collection has completed, those of data have been put into the ANFIS for training and validating so that we can get a trained ANFIS system for infer the key parameters such as the frame delay for video, the least volume in dB for voice receiver, and the size of audio buffer. After that, the video/voice over IP has been tested in the cloud computing system based on Web interface and as a result it performs very well on video phone call over IP.

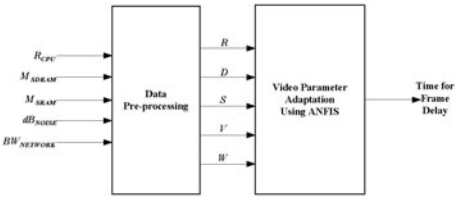


Fig. 12. Intelligent adaptation of video parameter

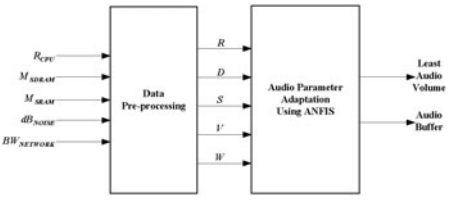


Fig. 13. Intelligent adaptation of audio parameter

4 Experimental Results and Discussions

A client/server scheme of VVoIP application running on Hadoop cloud computing will be denoted CLC-IHU in the following experiments, as shown in Figs. 14, 15, and 16, and discussions. Two remarkable benchmarks for the performance comparison of access security are revealed in FACE ID2 [36] and ZKS-F20 [37] where Equal Error Rate (EER), for the processes on facial recognition and fingerprint verification, and Response Time are two most concerned measures in the control of access security. As listed in Table 1, the comparison of performance with three models, FACE ID2, ZKS-F20, and CLC-IHU, is consequently shown that the method we proposed here outperforms the other alternatives due to fast response and low misclassification rate in access security.

In comparison to the previous work as shown in Fig. 17, the result of the experiment on VVoIP at Hadoop cloud computing, as shown in Figs. 18 and 19, is really to highly improve the network traffic for video/audio streaming, and easy-to-use TCP/IP protocol instead of PR-SCTP. A well-known benchmark for QoS measures is USHA scheme [38] where low latency and low packet loss are two most critical design issues in USHA scheme. As a result, the comparison of QoS with three models, USHA, SCTP-IHU, and CLC-IHU, as listed in Table 2 is shown that the

approach we proposed here is good enough to realize a real-time on-line video phone call between embedded platforms. In addition, we also emphasizes that real-time streaming multimedia with cloud computing scheme to achieve the least losses of transmitted packets, below 2% on average for video streaming, which is acceptable for four types of handoff. This result show the head-of-line blocking and handoff interruption can be resolved somehow by fast re-connection and parallel computation. According to the specification for mini2440 [39], ARM-based embedded platform, it is noticed that the power consumption is really reduced dramatically in mobile device when we adopt cloud computing scheme, CLC-TCP, for a real-time on-line video phone call.

In the application of VVoIP running in Hadoop cloud computing, the handoff delay time can achieve less than 1.7 sec as shown in Table 2. In order to verify the effectiveness and efficiency in access control for preventing illegal intrusions from the outside of the cloud, the rapid face recognition and fingerprint identification in Hadoop cloud computing has been done successfully within 2.2 seconds, as shown in Table 1, to exactly cross-examine the subject identity. As a result the proposed approach outperforms the other alternatives due to fast response and low misclassification rate like EER in access control.

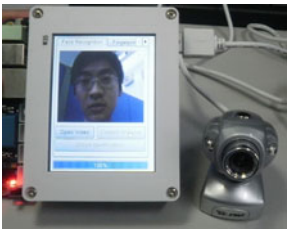


Fig. 14. Binarization processing automatically running in a program

Fig. 15. Processing fingerprint features to reduce the amount of information

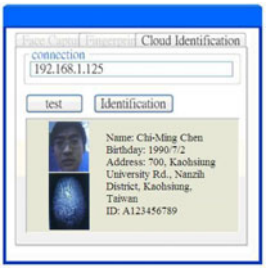


Fig. 16. Information sent to the cloud and cloud displays the results of recognition on the consol

Fig. 17. VVoIP over two embedded platform mini2440 via P2P connection

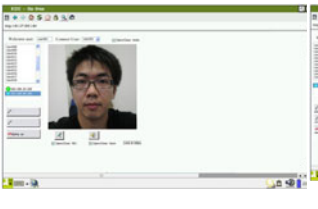


Fig. 18. CLC-IHU implementing VVoIP over two embedded platform mini2440 on user A

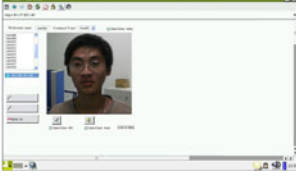


Fig. 19. CLC-IHU implementing VVoIP over two embedded platform mini2440 on user B

Table 1. The Performance Comparison of Access Security

Performance	FACE ID2	ZKS F20	CLC-IHU
Equal Error Rate (EER)	<0.1%	<0.01%	<0.01%
Response Time	<5.5 sec	<3.7 sec	<2.2 sec

Table 2. The QoS Comparison of VVoIP

Performance	USHA	SCTP-IHU	CLC-IHU
Handoff Delay	2.5 sec	1.865 sec	1.631 sec
Power Consumption at Thin Client	—	11.7mAV ∫ 10.53mAV	7.02mAV ∫ 6.318mAV

5 Conclusions

In the previous work, application programs for VVoIP have installed in thin client. In this study we have moved the VVoIP applications from thin client to a Hadoop cloud computing system where access control has been set up there using rapid facial and fingerprint identifications. User does not need to know what is a real IP and web interface achieves interaction by adopt TCP instead of PR-SCTP so taht CLC-IHU scheme reduces computation load and power consumption dramatically at thin clients. In addition, we employed ANFIS to tune key factors appropriately for adapting handoff and analyzing network traffic at any time. As a result it takes a very short delay for the seamless handoff between base stations under mobile wireless network. We finally draw the conclusion that the approach we proposed here achieves a better performance and efficiency than the previous works.

Acknowledgments. This work is supported by the National Science Council, Taiwan, Republic of China, under grant number NSC 99-2220-E-390-002.

References

1. RTP: A Transport Protocol for Real-Time Applications (RFC 3550 3551) (2003), http://www.ietf.org/rfc/rfc3550.txt
2. Chang, B.R., Young, C.P., Tsai, H.F., Fang, R.Y.: Timed PR-SCTP for Fast Voice/Video over IP in Wired/Wireless Environments. Journal of Information Hiding and Multimedia Signal Processing 2(4), 320–331 (2011)

3. Chang, B.R., Young, C.P., Tsai, H.F., Fang, R.Y.: Embedded System for Inter-Vehicle Heterogeneous-Wireless-based Real-Time Multimedia Streaming and Video/Voice over IP. In: 4th International Conference on Innovative Computing, Information and Control (ICICIC 2009), Paper ID: B06-0, Ambassador Kaohsiung, Taiwan, December 7-9 (2009)
4. P2P, wikipedia (2011), http://en.wikipedia.org/wiki/Peer-to-peer
5. Stewart, R., Xie, Q., Morneault, K., Sharp, C., Schwarzbauer, H., Taylor, T., Rytina, I., Kalla, M., Zhang, L., Paxson, V.: Stream Control Transport Protocol. IETF RFC2960 (October 2000)
6. ARM Platform Technical Guide 2011, TI (2011), http://focus.ti.com/lit/sg/sprt596/sprt596.pdf
7. Perkins, C., Calhoun, P.: Authentication, Authorization, and Accounting (AAA). IETF RFC5637 (March 2005)
8. Welcome to Apache Hadoop (2010), http://hadoop.apache.org/
9. Java 2 Platform, Micro Edition, J2ME (2010), http://www.java.com/zh_TW/download/faq/whatis_j2me.xml
10. JamVM – A compact Java Virtual Machine (2010), http://jamvm.sourceforge.net/
11. GNU Classpath, GNU Classpath, Essential Libraries for Java (2010), http://www.gnu.org/software/classpath/
12. Roger Jang, J.S.: ANFIS: Adaptive Network-Based Fuzzy Inference System. IEEE Transactions on System, Man and Cybernetics 23(3), 665–685 (1993)
13. Google App Engine (2010), http://groups.google.com/group/google-appengine
14. Amazon Web Services, AWS (2010), http://aws.amazon.com/
15. Windows Azure-A Microsoft Solution to Cloud, (2010), http://tech.cipper.com/index.php/archives/332
16. IBM Cloud Compputing (2010), http://www.ibm.com/ibm/cloud/
17. Open Nebula (2010), http://www.opennebula.org/
18. Eucalyptus (2010), http://open.eucalyptus.com/
19. Welcome to Apache Hadoop (2010), http://hadoop.apache.org/
20. Sector/Sphere, National Center for Data Mining (2009), http://sector.sourceforge.net/
21. Voice over IP in a Cloud, IIS (2009), http://www.17freecall.com/, http://www.loip.com/
22. VoIP in cloud computing, Skpye (2010), http://www.chinalabs.com/html/chanyezhuanxing/2010/1214/41378.html
23. Hicloud with CaaS & StaaS, Chunghwa Telecom (2011), http://hicloud.hinet.net/
24. Innovative DigiTech-Enabled Applications & Services Institute (2011), http://www.ideas.iii.org.tw/index.html
25. NCHC Cloud Computing Research Group (2010), http://trac.nchc.org.tw/cloud
26. Cloud computing Center for Mobile Application, ITRI (2011), http://www.itri.org.tw/chi/ccma/
27. Hadoop4win, Hadoop for Windows (2011), http://sourceforge.net/projects/hadoop4win/files/0.1.3/hadoop4win-setup-full_0.1.3.zip/download\

28. Hadoop (2011), http://apache.stu.edu.tw//hadoop/core/hadoop-0.20.2/hadoop-0.20.2.tar.gz
29. Eclipse Summit (2011), http://www.eclipse.org/dowload/
30. Java JDK (2011), http://www.oracle.com/technetwork/java/javase/downloads/jdk6-jsp-136632.html
31. Qt/Embedded, Embedded Linux (2011), http://qt.nokia.com/products/platform/qt-for-embedded-linux/
32. Java Native Interface (JNI), Oracle (2011), http://java.sun.com/docs/books/jni/
33. Veri Finger SDK, Neuro Technology (2010), http://www.neurotechnology.com/verifinger.html
34. Veri Look SDK, Neuro Technology (2010), http://www.neurotechnology.com/verilook.html
35. Opencv, open source (2010), http://www.opencv.org.cn/index.php?title=%E9%A6%96%E9%A1%B5&variant=zh-tw
36. FACE ID2, fingertecusa (2011), http://fingertecusa.com/face-recognition-model-c-53/face-id-2-p-92
37. ZKS-iClock 7 & ZKS-F20, ZKS Group (2011), http://lancyzks.bossgoo.com/product-Video-Door-Phone/ZKS-iClock-7Fingerprint-Time-Attendance-Access-Control-933894.html, http://lancyzks.bossgoo.com/product-Video-Door-Phone/ZKS-F20-STANDALONE-FACE-RECOGNITION-ACCESS-SYSTEM-933839.html
38. Chen, L.J., Sun, T., Cheung, B., Nguyen, D., Gerla, M.: Universal Seamless Handoff Architecture in Wireless Overlay Networks. Technical Report TR040012, UCLA CSD (2004)
39. Marvell PXA3xx (88AP3xx) Processor Family (2001), http://www.marvell.com/products/processors/applications/pxa_family/PXA3xx_EMTS.pdf

Intelligence Decision Trading Systems for Stock Index

Monruthai Radeerom[1], Hataitep Wongsuwarn[2], and M.L. Kulthon Kasemsan[1]

[1] Faculty of Information Technology, Rangsit University,
Pathumtani, Thailand 12000
[2] ME, Faculty of Engineering, Kasetsart University (Kamphaeng Saen),
Nakhonpathom, Thailand 73140
mradeerom@yahoo.com, fenghtw@ku.ac.th,
kasemsan@rangsit.rsu.ac.th

Abstract. This paper introduces an intelligent decision-making model, based on the application of Fuzzy Logic and Neurofuzzy system (NFs) technology. Our proposed system can decide a trading strategy for each day and produce a high profit for each stock. Our decision-making model is used to capture the knowledge in technical indicators for making decisions such as buy, hold and sell. Moreover, we compared with 3 our proposed scenario of Intelligence Trading System model. Finally, the experimental results have shown higher profits than the Neural Network (NN) and "Buy & Hold" models for each stock index. And, some models which were including volume indicator and predicted close price on next day have profit batter than other models. The results are very encouraging and can be implemented in a Decision- Trading System during the trading day.

Keywords: Computational intelligence, Neuro-Fuzzy System, Stock Index, Decision Making System.

1 Introduction

The prediction of financial market indicators is a topic of considerably practical interest and, if successful, may involve substantial pecuniary rewards. People tend to invest in equity because of its high returns over time. Considerable efforts have been put into the investigation of stock markets. The main objective of the researchers is to create a tool, which could be used for the prediction of stock markets fluctuations; the main motivation for this is financial gain. In the financial marketplace, traders have to be fast and hence the need for powerful tools for decision making in order to work efficiently, and most importantly, to generate profit.

The use of Artificial Intelligence (AI) had a big influence on the forecasting and investment decision-making technologies. There are a number of examples using neural networks in equity market applications, which include forecasting the value of a stock index [4,5], recognition of patterns in trading charts[12,17] rating of corporate bonds[8], estimation of the market price of options[11], and the indication of trading signals of selling and buying[3,7].

J.-S. Pan, S.-M. Chen, N.T. Nguyen (Eds.): ACIIDS 2012, Part III, LNAI 7198, pp. 366–375, 2012.

Even though most people agree on the complex and nonlinear nature of economic systems, there is skepticism as to whether new approaches to nonlinear modeling, such as neural networks, can improve economic and financial forecasts. Some researchers claim that neural networks may not offer any major improvement over conventional linear forecasting approaches [8, 12]. In addition, there is a great variety of neural computing paradigms, involving various architectures, learning rates, etc., and hence, precise and informative comparisons may be difficult to make. In recent years, an increasing amount of research in the emerging and promising field of financial engineering has been incorporating Neurofuzzy approaches [10, 12, 16,19,20]. Almost all models are focused on the prediction of stock prices. The difference of our proposed model is that we are focusing on decision-making in stock markets, but not on forecasting in stock markets.

In contrast to our previous work [15], we are not making a direct prediction of stock markets, but we are working on a one-day forward decision-making tool for buying/selling stocks. We are developing a decision-making model which works beyond the application of Fuzzy Logic and Neuro-Fuzzy systems (NFs). At first, our proposed trading strategy based on Fuzzy Logic captured knowledge from experts who are making decisions to buy, hold, or sell from technical analysis as well as input from our proposed trading systems based on NFs. Moreover, optimization algorithms based on the rate of the return profit of each stock index constructed our NFs model. In this paper, we present a decision-making model which combines technical analysis and NFs models. The technical analysis model evaluated knowledge about buy, hold and sell strategies from each technique. Our proposed model used results from the technical analysis model to input into our NFs. The NFs trading system decides the buy, sell and hold strategy for each stock index. The objective of this model is to analyze the stock daily and to make one day forward decisions related to the purchase of stocks.

The paper is organized as follows: Section 2 presents the background about the neural network and the Neurofuzzy system. Section 3 presents the NFs decision-making model; Sections 4 is devoted to experimental investigations and the evaluation of the decision-making model. This section provides the basis for the selection of different variables used in the model, and models the structure. The main conclusions of the work are presented in Section 5, with remarks on future directions.

2 Neuro-Fuzzy Approaches for the Intelligence Decision Trading System

Both neural networks and the fuzzy system imitate human reasoning process. In fuzzy systems, relationships are represented explicitly in forms of if-then rules. In neural networks, the relations are not explicitly given, but are coded in designed networks and parameters. Neurofuzzy systems combine the semantic transparency of rule-based fuzzy systems with the learning capability of neural networks. Depending on the structure of if-then rules, two main types of fuzzy models are distinguished as mamdani (or linguistic) and takagi-sugeno models [18]. The mamdani model is typically used in knowledge-based (expert) systems, while the takagi-sugeno model is used in data-driven systems

In this paper, we consider only the Takagi - Sugeno model. Takagi and Sugeno [18] formalized a systematic approach for generating fuzzy rules from an input-output

data pairs. The fuzzy if-then rules, for the pure fuzzy inference system, are of the following form:

$$if\ x_1\ is\ A_1\ and\ x_2\ is\ A_2\ and\ x_N\ is\ A_N\ then\ y = f(x) \tag{1}$$

Where $x = [x_1, x_2, ..., x_N]^T$, $A_1, A_2, \ldots, A_N$ fuzzy sets are in the antecedent, while y is a crisp function in the consequent part. The function is a polynomial function of input variables $x_1, x_2, x_3, \ldots, x_N$.

The first-order TSK fuzzy model could be expressed in a similar fashion. Consider an example with two rules:

$$\begin{aligned} &if\ x_1\ is\ A_{11}\ and\ x_2\ is\ A_{21}\ and\ then\ y_1 = p_{11}x_1 + p_{12}x_2 + p_{10} \\ &if\ x_1\ is\ A_{12}\ and\ x_2\ is\ A_{22}\ and\ then\ y_2 = p_{21}x_1 + p_{22}x_2 + p_{20} \end{aligned} \tag{2}$$

Fig. 1 shows a network representation of those two rules. The nodes in the first layer compute the membership degree of the inputs in the antecedent fuzzy sets. The product node $\prod$ in the second layer represent the antecedent connective (here the "and" operator). The normalization node N and the summation node $\sum$ realize the fuzzy-mean operator for which the corresponding network is given in Fig. 1. Applying fuzzy singleton, a generalized bell function such as membership function and algebraic product aggregation of input variables, at the existence of M rules the Neurofuzzy TSK system output signal upon excitation by the vector, are described by

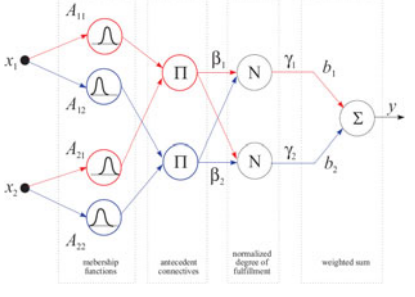


Fig. 1. An example of a first-order TS fuzzy model with two rules systems [1]

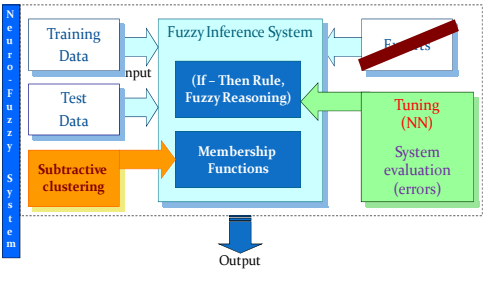


Fig. 2. Constructing Neurofuzzy Networks

In conclusion, Fig. 2 summarizes the Neurofuzzy Networks System (NFs). Construction process data called "training data sets," can be used to construct Neurofuzzy systems. We do not need prior knowledge ala "knowledge-based (expert) systems". In this way, the membership functions of input variables are designed by the subtractive clustering method. Fuzzy rules (including the associated parameters) are constructed from scratch by using numerical data. And the parameters of this model (the membership functions, consequent parameters) are then fine-tuned by process data.

3 Methodology for the Intelligence Decision Trading System

3.1 Decision-Making Model for the Stock Market System

Many stock market traders use conventional statistical techniques for decision-making in purchasing and selling [7]. Popular techniques use fundamental and technical analysis. They are more than capable of creating net profits within the stock market, but they require a lot of knowledge and experience. Because stock markets are affected by many highly interrelated economic, political and even psychological factors, and these factors interact with each other in a very complex manner, it is generally very difficult to forecast the movements of stock markets (see Fig 3). Fig. 3 shows historical quotes of Bangchak Petroleum Public Co., Ltd. (BCP) stock prices. It is a high nonlinear system. In this paper, we are working on one day decision making for buying/selling stocks. For that we are developing a decision-making model, besides the application of an intelligence system. We selected a Neurofuzzy system (NFs), which are now studied and incorporated into the emerging and promising field of financial engineering [2, 8, 10, 14].

We proposed NFs for our decision-making model, which we call Intelligence Trading System. The model scenario represents one time calculations made in order to produce decisions concerning the trading of stocks. For this paper, historical data of daily stock returns was used for the time interval. In the first step of the model realization, technical analysis techniques are used for the decision strategy recommendation. The recommendations (R) represent the relative rank of investment attraction to each stock in the interval [−1, 1]. The values −1, 0, and 1 represent recommendations: Sell, Hold and Buy, respectively. After that, the recommendations are included in the input of our proposed intelligence system. The intelligence system output is the evaluating recommendation based on several decided courses of action from various technical techniques used by investors as shown on Fig. 4.

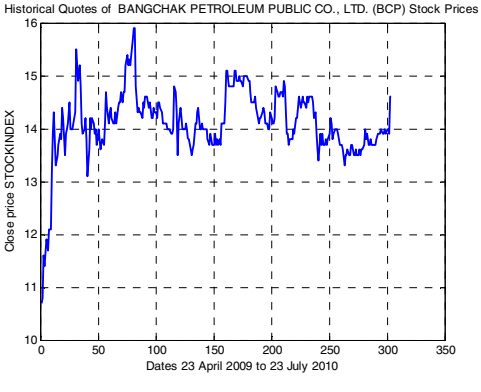


Fig. 3. Historical Quotes of Bangchak Petroleum public Co., Ltd. (BCP) Stock Prices

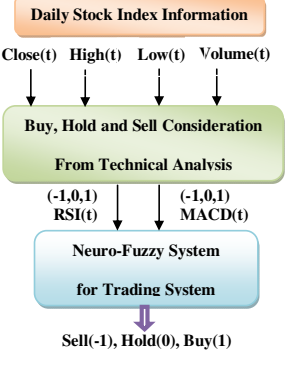


Fig. 4. The scenario of Intelligence Trading System

3.2 Input Variables Based on Technical Analysis and Evaluating Profit

Technical analysts usually use indicators to predict future buy and sell signals. The major types of indicators are Moving Average Convergence/Divergence (MACD), Williams's %R (W), Relative Strength Index (RSI), Exponential Moving Average (EMA), On Balance Volume, etc., and they correspond on close price and volume. These indicators can be derived from the real stock composite index. Each indicator is included in the input signal for the intelligence system. And, the target for training is the buy and sell signal as shown on Fig 5.

For daily data, indicators can help traders identify trends and turning points. The moving average is a popular and simple indicator for trends. Stochastic and RSI are some simple indicators which help traders identify turning points. Some example indicators such as Relative Strength Index (RSI) are defined as follows,

$$RSI = 100 - \frac{100}{1 + \frac{\sum(positive\ change)}{\sum(negative\ change)}} \tag{3}$$

Fig. 5 (a) shown a RSI index in 14 periods time. After that, we can be evaluating buy, hold and sell signal by experience trader. Sell zone was RSI more than 70 and buy zone was RSI below 30. Thus, Hold zone was between 30 and 70. Sell, buy and hold zone shown Fig 5(b).

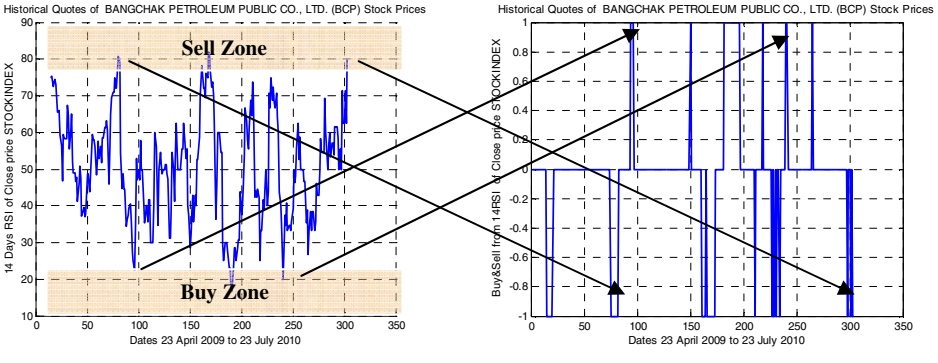


Fig. 5. (a) The 14 periods RSI index (RSI14 (t)) calculated by close price(t)
(b) The Buy (1), Hold (0), Sell (-1) evaluated by RSI14(t)

Because the index prices and technical indicators are not in the same scale, the same maximum and minimum data are used to normalize them. Normalization can be used to reduce the range of the data set to values appropriate for inputs to Neuro-Fuzzy system. The max is derived from the maximum value of the linked time series; similarly minimum is derived from the minimum value of the linked time series. The maximum and minimum values are from the training, Testing and validation data sets

must be same scale. The outputs of the Neuro-Fuzzy will be rescaled back to the original value according to the same formula. After normalization, input scale is between -1 and 1. The normalization and scaling formula is

$$y = \frac{2x - (\max + \min)}{(\max - \min)}, \tag{4}$$

Where x is the data before normalizing, y is the data after normalizing.

For NFs trading system, the expected returns are calculated considering the stock market. That is, the value obtained on the last investigation day is considered the profit. The trader's profit is calculated as

$$\text{Profit(n)} = \text{Stock Value(n)} - \text{Investment value} \tag{5}$$

Where n is the number of trading days.

And the Rate of Return Profit (RoRP) is

$$\text{RoRP} = \frac{\text{Profit(n)}}{\text{Investment value}} \times 100 \tag{6}$$

4 Results and Discussion

The model realization could be run having different groups of stocks (like Banking group, Energy group, etc.), indexes or other groups of securities. For that, we are using market orders as it allows simulating buying stocks, when stock exchange nearly closed market. All the experimental investigations were run according to the above presented scenario and were focused on the estimation of Rate of Return Profit (RoRP). At the beginning of each realization, the start investment is assumed to be 1,000,000 Baht (Approximately USD 29,412). The data set, including the Stock Exchange of Thailand (SET) index, Historical Quotes of Bangchak Petroleum public Co., Ltd. (BCP) Stock Prices, Siam Commercial Bank (SCB) and Petroleum Authority of Thailand (PTT) stock index, has been divided into two different sets: the training data and test data. The stock index data is from April 23, 2009 to July 23, 2010 totaling 304 records. The first 274 records are training data, and the rest of the data, i.e., 30 records, will be test data. Moreover, the data for stock prices includes the buy-sell strategy, closing price and its technical data. Consequently, max-min normalization can be used to reduce the range of the data set to appropriate values for inputs and output used in the training and testing method.

4.1 Input Variables

Technical indexes are calculated from the variation of stock price, trading volumes and time according to a set of formulas to reflect the current tendency of the stock price fluctuations. These indexes can be applied for decision making in evaluating the phenomena of oversold or overbought stock. For input data, several technical indexes are described as shown in Table 1. We proposed 3 input models. Our 3 models are

difference number of inputs and 1 output which is trading signal (sell (-1), hold (0) and buy (1). Model 1 was totally 7 inputs. Model 2 was totally12 inputs. Model 3 was totally12 inputs.

Table 1. Input and output of NFs for Decision Trading System

Model 1

NO.	DESCRIPTION
	INPUT
1	Close Price (t)
2	Buy & Sell from RSI 4 Days (t)
3	Buy & Sell from RSI 9 Days (t)
4	Buy & Sell from 14 Days(t)
5	Buy & Sell from William 10 Days (t)
6	Buy & Sell from MCAD 10 Days (t)
7	Buy & Sell from EMA 10 and 25 Days (t)
	OUTPUT
1	Buy (1), Hold(2) and Sell(-1)

Model 2

NO.	DESCRIPTION
	INPUT
1	Close Price (t)
2	Typical Price (t)
3	Volume Rate of Change (t)
4	Price and Volume Trend (t)
5	Price Rate of Change 12 Days (t)
6	On-Balance Volume(t)
7	Buy & Sell from RSI 4 Days (t)
8	Buy & Sell from RSI 9 Days (t)
9	Buy & Sell from 14 Days(t)
10	Buy & Sell from William 10 Days (t)
11	Buy & Sell from MCAD 10 Days (t)
12	Buy & Sell from EMA 10 and 25 Days (t)
	OUTPUT
1	Buy (1), Hold(2) and Sell(-1)

Model 3

NO.	DESCRIPTION
	INPUT
1	Up and Down of Close Price (t+1)
2	Typical Price (t)
3	Volume Rate of Change (t)
4	Price and Volume Trend (t)
5	Price Rate of Change 12 Days (t)
6	On-Balance Volume(t)
7	Buy & Sell from RSI 4 Days (t)
8	Buy & Sell from RSI 9 Days (t)
9	Buy & Sell from 14 Days(t)
10	Buy & Sell from William 10 Days (t)
11	Buy & Sell from MCAD 10 Days (t)
12	Buy & Sell from EMA 10 and 25 Days (t)
	OUTPUT
1	Buy (1), Hold(2) and Sell(-1)

Moreover, we proposed 3 scenario models as shown on Fig. 6. Model 1 used close price and decision values from technical analysis as input to NFs. Model 2 was difference from Model1 which are including some volume indicators as inputs of NFs. Model 3 was difference from Model 2 which are using NFs for predicting close price on next day as input of Decision Trading NFs.

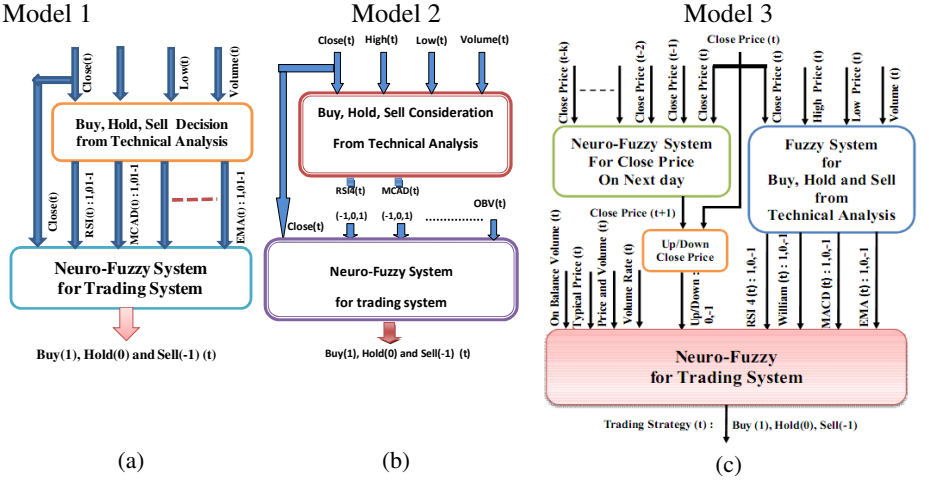


Fig. 6. Our proposed scenarios of Intelligence Decision Trading System

4.2 Evaluating Decision-Making System Based on Neurofuzzy Model

We now compare the performance of our proposed NFs with NNs including three types of learning algorithm methods. The learning methods are Batch Gradient

Descent (TRAINGD), Scaled Conjugate Gradient (TRAINSCG) and Levenberg-Marquardt (TRAINLM) methods The neural network model has one hidden layer with 30 nodes. And, learning iteration is 10000 epochs. After we trained their learning method, we found scaled conjugate better than other learning methods. Actually, we can conclude that our proposed Neurofuzzy demonstrated four relation types considerably better than NNs with scaled conjugate gradient learning.

After developing the intelligence trading system, we were given 1,000,000 baht for investment at the beginning of the testing period. The decision for to buy and sell stocks is given by proposed intelligence output. We translated the produced RoRP results that verify the effectiveness of the trading system. Table 2 is show a Financial Simulation Model for calculating profit in our trading strategy. For example of experimental result, results of Model 3 within training days and testing day are shown in Figure 7 and 8, respectively.

Table 2. Example of Financial Simulation Model in trading strategy

Stock	Possible Buy&Hold				NFs Buy & Sell			
Index	Action		# of Shared	Cash(Baht)	Action		# of Shared	Cash(Baht)
12.00	1	STAY	-	1,000,000.00	1	STAY	-	1,000,000.00
12.00	0	HOLD	-	1,000,000.00	1	HOLD	-	1,000,000.00
12.00	1	BUY	83,333		0	HOLD	-	1,000,000.00
12.50	0	HOLD	83,333		1	BUY	80,000	-
13.00	-1	SELL	-	1,083,329.00	0	HOLD	80,000	
12.70	0	HOLD	-	1,083,329.00	0	HOLD	80,000	
12.60	0	HOLD	-	1,083,329.00	0	HOLD	80,000	
12.55	0	HOLD	-	1,083,329.00	-1	SELL	-	1,004,000.00
12.45	0	HOLD	-	1,083,329.00	0	HOLD	-	1,004,000.00
12.40	1	BUY	87,365		0	HOLD	-	1,004,000.00
12.45	-1	SELL	-	1,087,694.25	-1	SELL	-	1,004,000.00
.....	...	.....	.....	.....	...	.....	.....	.....

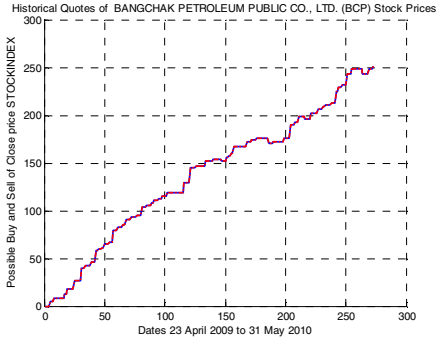


Fig. 7. Comparison profit between Possible Rate of Return Profit (Possible RoRP) and Profit from our proposed NFs Trading System in Training Days

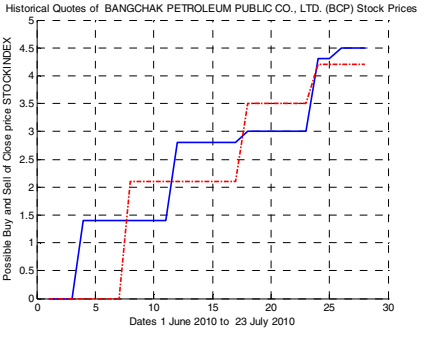


Fig. 8. Comparison profit between Possible Rate of Return Profit (Possible RoRP) and Profit from our proposed NFs Trading System in Testing Days

Moreover, our proposed decision-making NFs model compared RoRP performance with Buy & Hold Strategy and NNs. The antithesis of buy and hold is the concept of day trading in which money can be made in the short term if an individual tries to short on the peaks, and buy on the lows with greater money coming with greater volatility.

The performance by each stock index is illustrated on Table 3. It reflects the performances of investment strategies in NN, buy and hold and NFs model, respectively. Each line implies the performance of the NFs system in terms of cumulative profit rate of return gained from each stock index. In the case of experimental results, NFs display a greater rate of return than the "buy, sell and hold" model and NN model. Moreover, when comparing of 3 Models, Profit from Model 1 is lower than Model 2 and 3. Differencing Model 1 from Model 2 and 3 was not volume indicator. Model 3 was profit than Model 2 because it was including close price next day into its model.

The results of difference in the stock index results are small. It is more valuable to calculate the loss and gains in terms of profitability in practice.

Table 3. Rate of Return Profit (RoRP) gained from each trading stock index

Stock Index	Stock Group	Possible Profit	Profit Buy & Hold	Model 1.		Model 2.		Model 3.	
		Training Days(%)		Profits of Training Days (274 Days). %					
				NN	NFs	NN	NFs	NN	NFs
BCP	Energy	254	50	90	115	240	250	240	254
PTT	Energy	320	80	115	180	300	308	300	320
SCB	Banking	180	70	75	98	160	172	160	180
Stock Index	Stock Group	Testing Days(%)		Profits of Testing Days (28 Days). %					
				NN	NFs	NN	NFs	NN	NFs
BCP	Energy	4.5	0.8	1.2	1.4	3.3	4.1	4.5	4.3
PTT	Energy	7.1	1.5	2.5	3.6	4.5	5.9	7.1	6.9
SCB	Banking	3.2	0.5	1.1	2.1	2.1	2.5	3.2	2.9

5 Conclusion

This paper presented the decision-making model based on the application of NFs. The model was applied in order to make a one-step forward decision, considering from historical data of daily stock returns. The experimental investigation has shown scenario of Intelligence Trading System based on Fuzzy and NFs to make a trading strategy, achieving more stable results and higher profits when compared with NNs and Buy and Hold strategy. Moreover, we compared with 3 our proposed scenario of Intelligence Trading System model. And, some models which were including volume indicator and predicted close price on next day have profit batter than other models. For future work, several issues could be considered. Other techniques, such as support vector machines, genetic algorithms, etc. can be applied for further comparisons. And other stock index groups, another stock exchange or other industries in addition to electronics one is further considered for comparisons.

Acknowledgment. This piece of work was partly under the Graduate Fund for Ph.D. Student, Rangsit University, Pathumthanee, Thailand.

References

1. Babuska, A.R.: Neuro-fuzzy methods for modeling and identification. In: Recent Advances in Intelligent Paradigms and Application, pp. 161–186 (2002)
2. Cardon, O., Herrera, F., Villar, P.: Analysis and Guidelines to Obtain A Good Uniform Fuzzy rule Based System Using simulated Annealing. J. Approximated Reason. 25(3), 187–215 (2000)
3. Chapman, A.J.: Stock market reading systems through neural networks: developing a model. J. Apply Expert Systems 2(2), 88–100 (1994)
4. Chen, A.S., Leuny, M.T., Daoun, H.: Application of Neural Networks to an Emerging Financial Market: Forecasting and Trading The Taiwan Stock Index. Computers and Operations Research 30, 901–902 (2003)
5. Conner, N.O., Madden, M.: A Neural Network Approach to Prediction Stock Exchange Movements Using External Factor. Knowledge Based System 19, 371–378 (2006)
6. Liu, J.N.K., Kwong, R.W.M.: Automatic Extraction and Identification of chart Patterns Towards Financial Forecast. Applied Soft Computing 1, 1–12 (2006)
7. Doeksen, B., Abraham, A., Thomas, J., Paprzycki, M.: Real Stock Trading Using Soft Computing Models. In: IEEE Int'l Conf. on Information Technology: Coding and Computing, Las Vegas, Nevada, USA, pp. 123–129 (2005)
8. Dutta, S., Shekhar, S.: Bond rating: A non-conservative application of neural networks. In: IEEE Int'l Conf. on Neural Networks, San Diego, CA, USA, pp. 124–130 (1990)
9. Farber, J.D., Sidorowich, J.J.: Can new approaches to nonlinear modeling improve economic forecasts? The Economy As An Evolving Complex System, 99–115 (1988)
10. Hiemstra, Y.: Modeling Structured Nonlinear Knowledge to Predict Stock Markets: Theory. In: Evidena and Applications, Irwin, pp. 163–175 (1995)
11. Hutchinson, J.M., Lo, A., Poggio, T.: A nonparametric approach to pricing and hedging derivative securities via learning networks. J. Finance 49, 851–889 (1994)
12. James, N.K., Raymond, W.M., Wong, K.: Automatic Extraction and Identification of chart Patterns towards Financial Forecast. Applied Soft Computing 1, 1–12 (2006)
13. LeBaron, B., Weigend, A.S.: Evaluating neural network predictors by bootstrapping. In: Int'l Conf. on Neural Information Process., Seoul, Korea, pp. 1207–1212 (1994)
14. Li, R.-J., Xiong, Z.-B.: Forecasting Stock Market with Fuzzy Neural Network. In: 4th Int'l Conf. on Machine Learning and Cybernetics, Guangzhou, China, pp. 3475–3479 (2005)
15. Radeerom, M., Srisaan, C.K., Kasemsan, M.L.K.: Prediction Method for Real Thai Stock Index Based on Neurofuzzy Approach. In: Trends in Intelligent Systems and Computer Engineering. LNEE, vol. 6, pp. 327–347 (2008)
16. Refenes, P., Abu-Mustafa, Y., Moody, J.E., Weigend, A.S. (eds.): Neural Networks in Financial Engineering. World Scientific, Singapore (1996)
17. Tanigawa, T., Kamijo, K.: Stock price pattern matching system: dynamic programming neural network approach. In: Int'l Conf. on Neural Networks, vol. 2, pp. 59–69 (1992)
18. Takagi, T., Sugeno, M.: Fuzzy identification of systems and its application to modeling and control. IEEE Trans. on System, Man and Cybernetics 5, 116–132 (1985)
19. Trippi, R., Lee, K.: Artificial Intelligence in Finance & Investing, Chicago, IL, Irwin (1996)
20. Tsaih, R., Hsn, V.R., Lai, C.C.: Forecasting S&P500 Stock Index Future with A Hybrid AI System. Decision Support Systems 23, 161–174 (1998)

Anomaly Detection System Based on Service Oriented Architecture

Grzegorz Kołaczek and Agnieszka Prusiewicz

Institute of Computer Science, Wrocław University of Technology, Poland
{grzegorz.kolaczek,agnieszka.prusiewicz}@pwr.wroc.pl

Abstract. The problem of the network security has been taken up since eighties and has been developed up to present day. A major problem of an automatic intrusion detection is that, it is difficult to make a difference between a normal and an abnormal user behaviour. We propose the framework of a distributed anomaly detection system based on Service Oriented Architecture (SOA). The main idea of SOA is to treat applications, systems and processes as encapsulated components, which are called services. These services are represented by input and output parameters and the semantic description of their functionalities. We assume that all the functionalities of our system are delivered by the Web services.

1 Introduction

Today's business feature is distribution and heterogeneity of applications that deliver diverse functionalities and based on different business models. Problems with gaining application interoperability and extensibility lead to searching other information technology solutions. A new architectural approach is called Service Oriented Architecture (SOA). SOA is a new way of thinking of business and designing applications. The main idea of SOA is to treat applications, systems and processes as encapsulated and reusable components, which are called services. These services are represented by input and output parameters and the semantic description of its functionalities [7,8]. Basic principles of SOA are as follows: 1) applications must make their functionalities available to other applications as services that might be combined as composite services, 2) interoperability must be gained regardless of technology, 3) orchestration of the business processes across suppliers, partners and customers must be possible.

SOA requires also new approaches to security. Security must be perceived as a service having all the features as the others services in the systems based on SOA. Hence security service is defined at the right level of granularity, from the point of view of a service consumer, technology- and context-independent. What is important is that functional aspects of security such as: authentication, data confidentiality, data integrity, protection against attacks, privacy and anomaly detection are standard and the same both for the systems based on monolithic and service oriented architectures [2].

J.-S. Pan, S.-M. Chen, N.T. Nguyen (Eds.): ACIIDS 2012, Part III, LNAI 7198, pp. 376–385, 2012.

In this work we propose the framework of an anomaly detection system based on Service Oriented Architecture (SOA). According to the SOA paradigm we treat a security as a service that is delivered on different levels of granulation. The main advantage of applying SOA paradigm into security area and delivery of all the functionalities of IDS system as Web services is that implemented algorithms and methods may be very easily exchanged according to the characteristics of the monitored network and requirements with respect to: 1) the level of security that must be maintained and 2) data mining methods used to obtain the profiles of the network system usage.

The systems for security evaluation based on SOA paradigm have been proposed in [10] where the definition of SOA security requirements for security evaluation process has been defined. The idea of service-based systems security level evaluation has been built on assumption that the security level is related to a few basic characteristics, e.g. profiles of services execution requests generated by system users and other services and the way how these requests are handled, the way of the complex services composition and execution or the utilization profile of the system resources. What is new here is the way of security evaluation that is carried out on different levels of granulation which include among others: the level of user, service, operating system or whole the system. In other words we evaluate here: 1) the security from the point of view of: 1) Web services that are used to deliver some functionalities, 2) users that use Web services and 3) a whole computer system.

We determine user and service profiles that are the basis for anomaly detection procedures. These profiles are aggregated characteristics of Web services usage.

2 The Functionalities and Architecture of Anomaly Detection System

The anomaly detection system proposed in this paper is a tool for monitoring the status of the system's security and identification of risk factors which can be used to improve the security management of the organization. It offers two classes of services. The first one is a monitoring service which examines selected characteristics of the system (e.g. patterns of communication between the services, the intensity of the communication, the time of implementation services, etc.) The second class of the provided services analyses the results obtained from the monitoring service and detects and identifies security incidents in heterogeneous and distributed service-oriented systems. The system provides its functionality using dedicated software agents. There are three classes of agents: agents specialized in detection of the global anomalies in the system behaviour, agents specialized in an anomaly detection of the separate service behaviour, and the agents which detect anomalies in a WLAN clients' behaviour.

The functionalities of our system refer to three levels of granulation of security evaluation and detection:

- the level of anomalous behaviour of Web services,
- the level of user anomalous behaviour ,
- the level of computer system anomalous behaviour.

3 The System Architecture

The general architecture of the anomaly detection systems consists of three layers. Three layered architecture introduces the following types of agents: service security level monitoring agents, client behaviour security level monitoring agents, system behaviour security level monitoring agents (Fig.1).

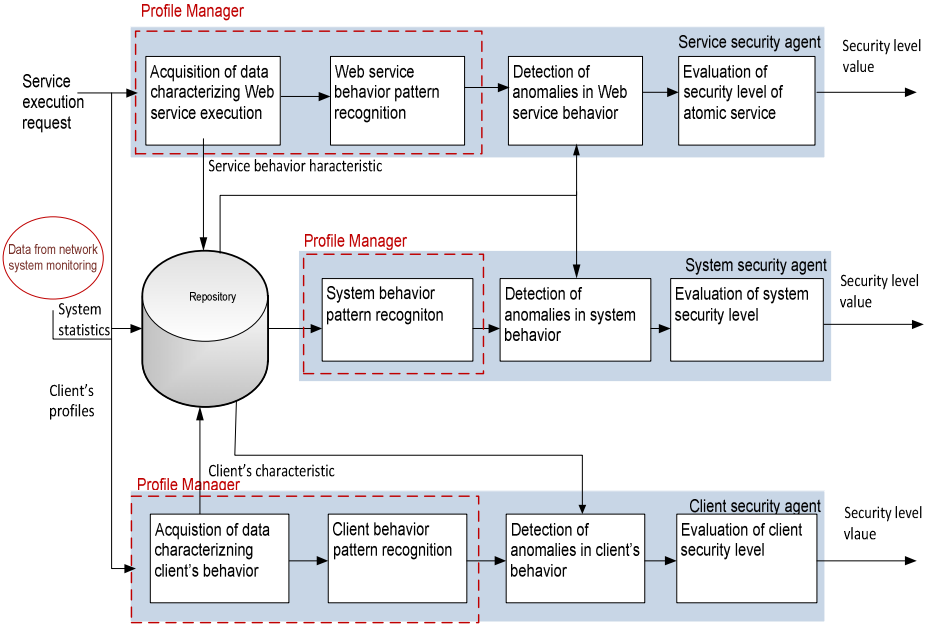


Fig. 1. The architecture of anomaly detection system

In the lowest layer of the mulit-agent system we get a set of specialized agents which are directly responsible for the monitoring and preliminary processing of data derived from the service-oriented system. This layer is the layer of the service security level agents.

The second layer – the agents responsible for the evaluation of security level of the system are used to aggregate the data from the agents of the lower layer. As a result, the level of the whole service-oriented system security can be evaluated.

The next layer is the layer of the agents which estimate the security level of the system clients. The task of the client security agent is to combine the data provided by lower layers agents and to automatically discover user behaviour patterns so to be able to detect their abnormal behaviour.

Some of security intrusions could be invisible or misinterpreted as the security analysis were limited only to information obtained from one from the above mentioned SOA layer. The complementary information from these three layers of SOA system allows to propose a unified method for the detection, correlation and prediction of security related events.

Profile Manager

One of the most crucial problems for intrusion detection or security evaluation is the ability to recognise abnormal states of the user, service or network system behaviour. We use Profile Manager for data acquisition and analysis in order to discover the pattern of behaviour on three levels of granularity: client, service and system. The advantage of applying SOA in our security system is the possibility of using different methods of knowledge engineering (implemented as Web services) to determine profiles of user behaviour or Web services usage. Hence the presented anomaly detection system is characterized by the following features and benefits: monitoring the activity of the service by using a dedicated interceptor activation, constant monitoring and identification of the communication patterns and generated network traffic between services, recording of the selected characteristics of the service activity to the local repository, detection of the abnormal service behaviour, evaluation and recording of the service security level in the local repository, detection of the abnormal system behaviour, evaluation and recording of the system security level in the local repository, abnormal behaviour detection of the clients of the service oriented system[11].

Profile Manager analyses data about the services usage and determines aggregated characteristics - user and service profiles, including extraction of patterns of activities, profiles of services and system usage. Profile Manager is composed of elements for clustering and classification of users and also pattern recognition (extraction), that are implemented as data mining Web services integrated by the Enterprise Service Bus (ESB). It is possible to exchange Web services with dependence on the method they implement by using the Profile Manager interface. Service Oriented Application Protocol (SOAP) is used for the communication between the Profile Manager interface and the data processing services [4].

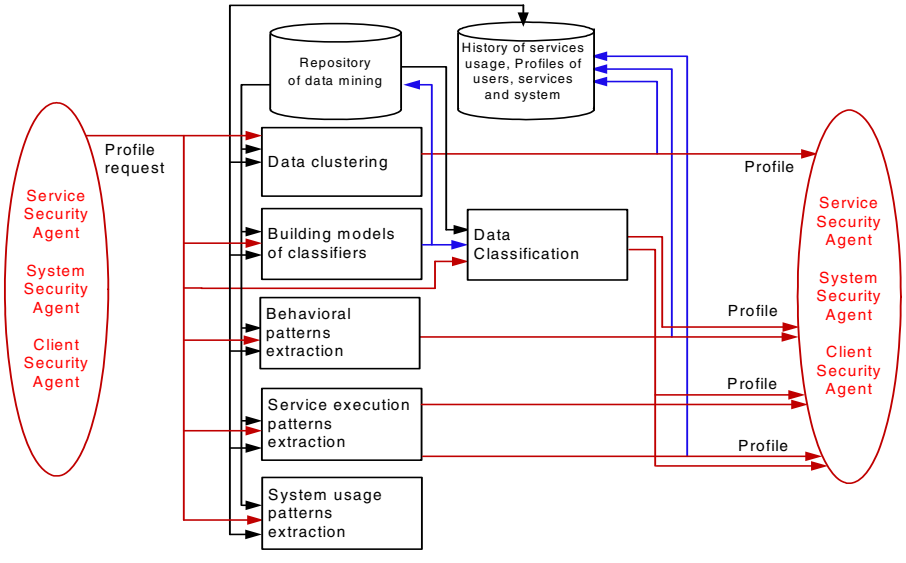


Fig. 2. Data Mining Services implemented in Profile Manager

The main goal of data processing is extraction of patterns according to service usage, system usage and client behaviour. Profile Manager is composed of the six functional blocks: data clustering, building classification models, data classification, behavioural pattern extraction, extraction patterns of service execution, system usage pattern extraction (Fig.3) [7].

The service security agent, system security agent and client security agent send to the Profile Manager the requests of profile determination. The profiles are obtained in functional blocks. The functionality hidden in data clustering block is responsible for grouping clients according to various criteria specified in clustering request. Using the block of building models of classifiers block it is possible to build models of classifiers using past observations, which contain vector of feature values of the objects and corresponding class labels. The built models of classifiers are used then in Data classification block to classify objects according to the vectors of feature values. Profile Manager has ability to select automatically the classification model by analyzing description of the data given in the classification request. Administrator can also chose classifiers from the list, which contains classifiers satisfying individual conditions about the classification accuracy [7]. The block of behavioural pattern extraction is responsible for determining aggregated characteristics of service usage, which contains statistical, temporal and sequence patterns of service usage. In the block System usage patterns extraction the general patterns of the clients's behaviors are determined. Then these patterns are used to discover any anomalies on the level of the service, the client and the system.

4 Profiles Determining

The key point in any anomalies detection is the knowledge about what we sanction as a normal. As a normal we take a determination the patterns of Web services usage, clients and system bahaviour. We use the Profile Manger to determine three classes of the profiles: Web service profile, client profile and system profile.

4.1 Web Service Profile

To determine $i-th$ Web service profile we measure the mean number of a Web service invocations at the time interval:

$$Service\,Pr\,ofile_{ij} = \left([t_b, t_e], n_{ij}\right)$$

where n_{ij} is the mean number of $i-th$ Web service invocations at the $j-th$ time interval $[t_b, t_e]$.Let us note that capturing only the mean number of the overall Web service invocations is not sufficient. Let us assume that profiles of four Web services: s_1, s_2, s_3, s_4 are given in the Table 1. However the mean numbers of all services usage are identical, characteristics of the services invocations at the particular time intervals are completely different.

Table 1. The profiles of the Web services for three time intervals

$[t_b,t_e]$	s_1	s_2	s_3	s_4
$[t_1,t_2]$	80	100	50	90
$[t_3,t_4]$	80	110	40	130
$[t_5,t_6]$	80	10	150	20
Mean value ($\bar{n}_{ij}$)	80	80	80	80

Hence taking into account the changeability of the Web service usage we represent Web service profile as a pair of: 1) time series of the mean number of services usage at the $j-th$ intervals and 2) the standard deviation of the number of the service usage computed for the obtained time series:

$$Service\,\mathrm{Pr}ofile_{ij}^{*} = \left(\frac{1}{2q+1} \sum_{r=-q}^{q} n_{ij,[t_b,t_e]+r,} \sigma_{n_{ij}} \right) \tag{1}$$

where:

– $^{n}ij_{[t_b,t_e]}$ is the component of the time series (the mean number of $i-th$ Web service invocations at the $j-th$ time interval),

– $\sigma_{n_{ij}} = \frac{1}{n_{ij}} \sqrt{\sum_{k=1}^{K} \left(n_{ijk} - \bar{n}_{ij} \right)^2}$

is the standard deviation of the number of a service invocation.

– $2q+1$ is the number of elements of a time series that are used to determine the moving average value.

Let us assume that at the time interval $[t_b,t_e]$ the number of the service s_i invocations is equal $Service_{i[t_b,t_e]}$.

$$AnomalyService_k = \begin{cases} true & if \quad \dfrac{\delta_{n_{ijk}}}{\sigma_{n_{ij}}} > 2\sqrt{2} \\ false & otherwise \end{cases} \tag{2}$$

where $\delta_{n_{ijk}} = \left| Service_{i[t_b,t_e]} - \bar{n}_{ij} \right|$

We use time series with the moving average value to reduce an accidental Web services usage and to smooth the values of $^{n}ij_{[t_b,t_e]}$.

4.2 User Profile

In turn a user profile at the time point t contains an aggregated knowledge (some tendencies) of the user services usage. We define the $k-th$ user profile as follows:

$$User\ Profile_{kt} = \{(s_i, (t_k, \Delta_k))\} \tag{3}$$

where:

- s_i denotes the Web service that is used by the $k-th$ user,
- (t_k, Δ_k) denotes a mean time of a day t with a Δ_k tolerance of the Web service invocations.

Let us assume that at the time interval $[t_b, t_e]$ the number of the service s_i usage by the $k-th$ user is equal $Service_{ikz\,t_b t_e}$.

To answer the question if at the time point t_z the anomaly of the $k-th$ user behaviour occurred we must apply the formulas (4):

$$AnomalyUser_k = \begin{cases} true\ if\ \left((s_i, (t_k, \Delta_k)) - Service_{ikz\,t_b t_e}\right) \geq \delta_1 \\ \quad and \left((s_i, (t_k, \Delta_k)) - Service_{ikz\,t_b t_e}\right) \leq \delta_2\ and \\ \quad [t_k - \Delta_k, t_k + \Delta_k] \subseteq [t_b, t_e] \\ false\quad otherwise \end{cases} \tag{4}$$

4.3 System Profile

The system profile has been defined for monitoring the system behaviour at the global level, which means that only changes in behaviour of multiple users and services (considered together) will be detected.

The system profile has been defined as a matrix (5) where columns are services and rows describe the services' behaviour related features extracted from the user and service profiles

$$System\,\mathrm{Pr}ofile_t = \begin{array}{l} \\ Userprofile_{1t}(t) \\ Userprofile_{1t}(\Delta t) \\ Userprofile_{2t}(t) \\ \cdots \\ Service\,\mathrm{Pr}ofile_t(\bar{n}_t) \\ Service\,\mathrm{Pr}ofile_t(\sigma_t) \end{array} \begin{array}{ccccccc} Service_{1t} & \ldots & Service_{nt} & Service_{n+1,t} & \ldots & Service_{m,t} \\ t_{11} & \ldots & t_{1n} & t_{1sn+1} & \cdots & t_{1sm} \\ \Delta t_{11} & \ldots & \Delta t_{1n} & \Delta t_{1sn+1} & \cdots & .\Delta t_{1sm} \\ t_{21} & \ldots & t_{2n} & t_{2sn+1} & \cdots & t_{1sm} \\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\ \bar{n}_{s1t} & \ldots & \bar{n}_{nt} & \bar{n}_{n+1t} & \ldots & \bar{n}_{mt} \\ \sigma_{s1t} & \ldots & \sigma_{nt} & \sigma_{n+1t} & \ldots & \sigma_{mt} \end{array} \tag{5}$$

where:

- $Userprofil\ e_{kt}(t)$ denotes the mean time of a day of the *k-th* Web service invocation
- $Userprofil\ e_{kt}(\Delta t)$ denotes Δ_k tolerance of the *k-th* Web service invocation
- $Service\,\mathrm{Pr}ofile_t(\bar{n}_t)$ denotes the mean number of $i-th$ Web service invocations at the $t-th$ time interval
- $Service\,\mathrm{Pr}ofile_t(\sigma_t)$ denotes the standard deviation of the number of service invocations
- $t_{11}, \Delta t_{11}, .., \bar{n}_{s1t}, \sigma_{s1t}$ are the values of above mentioned parameters at the $t-th$ time interval

The system's level anomaly detection is based on finding the principal eigenvector of the correlation matrix $System\,\mathrm{Pr}ofile_t$. The correlation matrix c is calculated, where c_{ij} means the correlation between services i and j taken from $System\,\mathrm{Pr}ofile_t$ for a fixed feature and fixed time window. When each feature is taken separately, we calculate M vectors for M features. By the principal vector for each feature we mean the vector which corresponds to the largest eigen value λ_{max}, which is the most informative vector and can be considered as a system's state. We detect anomaly when all principal vectors representing the current state change comparing to the previous state vectors for all the features we consider. This happens when multiple nodes change their behaviour. To determine the strength of the change we measure the distances between each of pair of the state vectors (current state vector and previous state vector). States are calculated in discreet moments of time and the interval length between those moments corresponds to the detector reaction time.

5 Detection of Security Related Events

The final opinion about the security level of the system under consideration generated by second's layer agents should take into account all three types of SOA system activity monitored by the first layer agents and reflect the measured properties of the services and users of services. However, in order to perform information fusion process integrating information from different layers a formal method which is required. Such a method, using Subjective Logic has been proposed in our earlier work [14] and due to space limitation will not be presented in details in this paper. In this case we need to express the information about service, user and system state in terms of Subjective Logic opinions. Opinions of Subjective Logic are tuples composed of three values ($\omega = \langle b,d,u \rangle$), for which the following interpretation have been defined:

1. *Belief* component of the Subjective Logic's opinion reflects the trust in the security level of the service under consideration. Its value close to one literally means that one perceives the service as safe.
2. *Disbelief* component reflects the opinion that a service is not safe and may cause security breaches.
3. *Uncertainty* component is to express that the knowledge is partial or incomplete and the security assessment does not give definite result.

It can be done by normalization of the anomaly values, e.g. for the Service we can assume that:

$$d = \min\{1, \frac{\bar{n} - n}{\sigma}\}, u = \min\{1, \frac{\sigma}{n}\}, b = 1 - d - u$$

Where d – disbelieve, u – uncertainty and b – believe component of the Subjective Logic opinion. Then generating the general opinion about the security level will a Subjective's Logic conjunction operator used to fuse opinions and resulting security assessment ω_{SS} for the system will have a form: $\omega_{SS} = \omega_{pbp} \wedge \omega_{srv} \wedge \omega_{srvd} \wedge \omega_{scp} \wedge \omega_{trl}$. According to the definition of conjunction operator, the belief component of the resulting opinion is close to the lowest security level measured for service layers – it is a case when the layers of the service architecture are treated as dependent from each other in the context of security assessment.

The second possible solution to global security assessment is just the application of the 'weakest link' paradigm. It means, that alert from any of the three considered layers will be equal to the security alert for the whole security system.

6 Conclusions

In this work the framework of the anomaly detection system based on service oriented architecture has been proposed. We presented the architecture and detailed functionality of this system. We introduced the procedure for behavioural pattern extraction. The patterns of users, services and system behaviour are then used by the agents to discover any anomalies in a network system. The functionalities related to the determination of the profiles of the system usage are delivered by the Web services. To discover the patterns of the user behaviours and applying the SOA approach to security system has a lot of advantages among others: reusability and exchangeability of the security services, interoperability with computer systems. The functionalities of our system are easily extensible. For further work it is necessary to conduct the tests of the functionalities of designed anomaly detection system in the chosen area.

Acknowledgements. The research presented in this work has been partially supported by the European Union within the European Regional Development Fund program no. POIG.01.03.01-00-008/08.

References

1. Juszczyszyn, K., Nguyen, N.T., Kolaczek, G., Grzech, A., Pieczynska, A., Katarzyniak, R.: Agent-Based Approach for Distributed Intrusion Detection System Design. In: Alexandrov, V.N., van Albada, G.D., Sloot, P.M.A., Dongarra, J. (eds.) ICCS 2006. LNCS, vol. 3993, pp. 224–231. Springer, Heidelberg (2006)
2. Kanneganti, R., Chodavarapu, P.: SOA Security. Manning Publications (2008)
3. Maselli, G., Deri, L., Suin, S.: Design and Implementation of an Anomaly Detection System: an Empirical Approach. In: Terena Networking Conference (TNC 2003), Zagreb, Croatia (May 2003)
4. Newcomer, E., Lomow, G.: Understanding SOA with Web Services. Addison Wesley Professional (2004)
5. Patcha, A., Park, J.-M.: An overview of anomaly detection techniques: Existing solutions and latest technological trends. Computer Networks 51(12), 3448–3470 (2007)
6. Prusiewicz, A.: On Some Method for Intrusion Detection Used by the Multi-Agent Monitoring System. In: Bubak, M., van Albada, G.D., Dongarra, J., Sloot, P.M.A. (eds.) ICCS 2008, Part III. LNCS, vol. 5103, pp. 614–623. Springer, Heidelberg (2008)
7. Prusiewicz, A., Zięba, M.: The Proposal of Service Oriented Data Mining System for Solving Real-Life Classification and Regression Problems. In: Camarinha-Matos, L.M. (ed.) Technological Innovation for Sustainability. IFIP AICT, vol. 349, pp. 83–90. Springer, Heidelberg (2011)
8. Rosen, M., Lublinsky, B., Smith, K.T., Balcer, M.J., Service-Oriented Architecture and Design Strategies. Wiley Publishing, Inc. (2008)
9. NIST/SEMATECH e-Handbook of Statistical Methods (2011), `http://www.itl.nist.gov/div898/handbook/`
10. Kołaczek, G.: Architecture for security level evaluation in service-based systems. In: Ruan, D. (ed.) Computational Intelligence: Foundations and Applications, pp. 844–850. World Scientific, New Jersey (2010)
11. Jagusiak, S., Kolaczek, G., et al.: Sniffer architecture for security level measurement in service oriented systems. In: Borzemski, L. (ed.) Information Systems Architecture and Technology: New Developments in Web-Age Information Systems, pp. 101–111. Oficyna Wydawnicza Politechniki Wrocławskiej, Wrocław (2010)
12. Kołaczek, G., Juszczyszyn, K.: Smart security assessment of composed web services. Cybern. Syst. 41(1), 46–61 (2010)
13. Juszczyszyn, K., Kołaczek, G., Prusiewicz, A.: Security assessment of composed Web services in a layered SOA security architecture. In: Ambroszkiewicz, S. (ed.) SOA Infrastructure Tools: Concepts and Methods, pp. 313–344. Poznań University of Economics Press, Poznań (2010)
14. Juszczyszyn, K., Kołaczek: Subjective logic-based framework for the evaluation of Web services' security. In: Grzech, A. (ed.) Information Systems Architecture and Technology: Service Oriented Distributed Systems: Concepts and Infrastructure, pp. 349–360. Oficyna Wydawnicza Politechniki Wrocławskiej, Wrocław (2009)

Efficient Data Update for Location-Based Recommendation Systems

Narin Jantaraprapa and Juggapong Natwichai

Computer Engineering Department, Faculty of Engineering
Chiang Mai University, Chiang Mai, Thailand
narin.jtp@gmail.com, juggapong@eng.cmu.ac.th

Abstract. Location-based recommendation systems are obtaining interests from the business and research communities. However, the efficiency of the update on the recommendation models is one of the most important issues. In this paper, we propose an efficient approach to update a recommendation model, User-centered collaborative location and activity filtering (UCLAF). The computational complexity of the model building is analyzed in details. Subsequently, our approach to update the models only the necessary parts is presented. As a result, the recommendation models obtained from our approach is exactly the same as the traditional re-calculation approach. The experiments have been conducted to evaluate our proposed approach. From the results, it is found that our proposed approach is highly efficient.

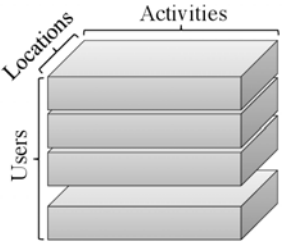
1 Introduction

Recommendation systems are playing an important role in the business these days. Not only the traditional systems, which recommend interesting items to the users, are utilized widely. But also the location-based recommendation systems, which can recommend places and activities to the users, are obtaining interests from the business and research communities. An example question which such system can answer would be "I want to go shopping, where should I go?". In this case, the answer can be the list of places with their ranking.

User-centered collaborative location and activity filtering (UCLAF) [6] is a prominent recommendation model which can recommend the information to the users accurately. It can recommend interesting locations and activities to users. When it is given the location, it can recommend the activities which can be attended nearby the location. If the activity, which a user wants to attend, is given, it can recommend the locations suited for such activity. The data model of UCLAF can be represented as 3D-matrix. In Figure 1a), an example data model is shown. It consists of three-axis i.e. users-axis, locations-axis and activities-axis. Each value in the matrix represent the frequency of a user attending an activity at a location.

If we consider only a single user, the 3D-matrix can be drill-downed to a 2D-matrix. For example in Figure 1b), the 2D-matrix of $user_1$ is shown. The matrix consists of the list of activities and the list of locations related to the user. As

J.-S. Pan, S.-M. Chen, N.T. Nguyen (Eds.): ACIIDS 2012, Part III, LNAI 7198, pp. 386–395, 2012.



a) UCLAF model

	$Activity_1$	$Activity_2$	$Activity_3$
$Location_1$	?	7	2
$Location_2$	7	?	?
$Location_3$	4	?	8

b) The 2D-matrix of $user_1$

	$Activity_1$	$Activity_2$	$Activity_3$
$Location_1$	1	7	2
$Location_2$	7	2	4
$Location_3$	4	3	8

c) The matrix of $user_1$

Fig. 1. UCLAF processes and examples

in the other recommendation models, the data can be very sparse due to variety of the users, the activities, and the locations. As it can be seen in Figure 1b), in which the question marks in the matrix represent N/A information for $user_1$ in the corresponding location. In order to recommend the information to $user_1$, the UCLAF algorithm, which is proposed with the model, can be applied to predict the missing frequency in Figure 1b) (the detail of the algorithm will be presented in Section 2). Figure 1c) shows the already-filled matrix by the algorithm. Subsequently, the model is ready for the recommendation. For example, suppose that $user_1$ wants to attend $activity_1$. From the model, the recommendation ranking is $location_2 \Rightarrow location_3 \Rightarrow location_1$. If $user_1$ wants an activity recommendation when he/she is located at $location_2$, the recommendation ranking of the activities is $activity_1 \Rightarrow activity_3 \Rightarrow activity_2$.

An important issue left in [6] is data updating. When users involve in many activities at many locations, the frequency values can change all the time. Some of the popular activities or locations in the past might not be that popular when the users request the recommendation. Thus, the algorithm needs to be re-applied to update the model. However, the computational expense of model computing is not small, i.e. $O(m^2nr) + O(mn^2r + n^3) + O(mn^2r + m^2n + n^3 + n^2p + nr^2) + O(mnr + mn^2)$, for user similarity computation, tensor decomposition computation, UCLAF algorithm computation, and tensor composition computation respectively. In which m is the number of users, n is the number of locations, r is the number of activities, and p is the number of features of the activities. In this paper, we propose an efficient approach to update the UCLAF model. In general, our approach begins with analyzing the computational complexity of the original UCLAF. Then, we utilize the analysis result to avoid the model re-calculation, when the updated data by the user are added, by calculation only the changed parts. Meanwhile the results of the model computing is

exactly the same. We have conducted the experiments to evaluate our proposed approach. A few parameter have been studied, i.e. memory usage and number of user-updates. From the experiment results, it is found that our proposed approach is highly efficient.

The organization of this paper is as follows. The related work is presented in the next section. Subsequently, our data updating approach is proposed in Section 3. In Section 4, the experiment results to evaluate our work are reported. Finally, Section 5 gives the conclusion and the outline of our future work.

2 Related Work

In this section, we present the related literatures which include recommendation systems and approaches to deal with the changed data.

For the recommendation systems, in [3], a survey on collaborative filtering methodologies such as item-based or user-based collaborative filtering as well as the implementation issues for such methods were presented. Another work is in [4], the authors applied the item-based collaborative filtering in their recommendation system. Such system represents the relation between the items and users preference in a 2D-matrix. Each value in the matrix represents the rating of a user on each item. In order to make a recommendation, the system requires three processing steps as follows. Firstly, item similarity is computed. Subsequently, the prediction computation is proceeded. This step applies an average weight method to compute, and predict a rating of user on each item. Finally, the ranking of items is recommended to the user.

In [7], a location-based recommendation system, Collaborative Location and Activity Recommendation (CLAR) model was proposed. The model, which can be presented as a 2D-matrix, consists of a list of locations and a list of activities. Each value in the matrix is adapted from GPS history data of a user who visited some places and did some activities. The matrix, when the data are collected, is not complete because some values can be missing. So, the authors proposed to fill the missing values by CLAR algorithm. The algorithm fills the matrix by augmenting two matrices into the input. The first matrix is a locations-features matrix, which each value represents the relation between a number of POIs, or feature, with each location. The other matrix is an activities-activities matrix, which each value in the matrix represents a correlation between activities.

For the changed data processing, in [2], the authors presented a system which can cope with the data changing. They introduced a data span approach with two window options, i.e. unrestricted window and the most recent window. Each option has its class of algorithms which can be used in different situations. The efficiency of the incremental processing is obtained from the data dividing into blocks with appropriate size.

In [1], the authors presented an algorithm to manage SVM (Support Vector Machine) models, which typically require large memory space as well as computational expense in the training phase. They proposed a new data incremental algorithm to update the SVM models. The algorithm partitions the SVM models into a number of blocks and stores such model into the memory. When a

new block of data is inserted, such data are incrementally used to train the recent models, and also build the new model. Subsequently, the set of models can classify the incoming test data. The characteristic of the models is stored in the latest blocks. The result of this new algorithm is very similar to the batch-version of it but can resolve the mentioned data update problems.

3 Our Solution and Complexity

In this section, we present our proposed approach to address the update problem. First, the original UCLAF is presented, its three-steps processing is illustrated in term of the computational complexity. Subsequently, our approach, which intends to cope with the scenarios that the users attend some activity at some places and update the data, is presented.

3.1 Original UCLAF

In order to recommend the information to the users, the UCLAF approach is proceeded as shown in Figure 2. It aims at lower the dimension since the data in the high-dimension can be too sparse to process. From the figure, there are three processing steps as follows. The first step is a tensor decomposition. Thus, the 3D-matrix represented the incomplete UCLAF model is decomposed into matrix X, Y and Z. Each matrix represents users, locations and activities. These matrices will be used in the UCLAF algorithm computation by augmenting the four matrices to them i.e. 1) Users-Users matrix 2) Location-Features matrix 3) Activities-Activities matrix, and 4) Users-Locations matrix. The Users-Users matrix represents the user similarity between users. The Location-Features matrix represents the relationship between the features and the locations. The Activities-Activities matrix represents the activity correlations. And, the Users-Locations matrix represents the frequency of the users visited the locations. When this step has completed, the new tensor will be composed from the new matrix X', Y', and Z'. Then, it can be used for the recommendation.

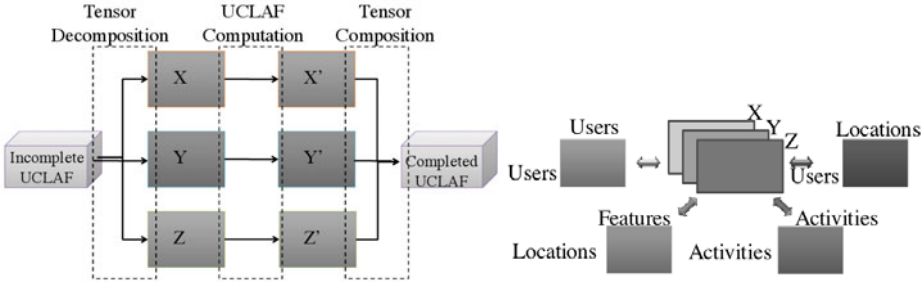


Fig. 2. Calculation process

3.2 Our Solution and Complexity

When a user attends an activity at a location, the models, i.e. the frequency values in the tensor matrix, the similarity values in the Users-Users matrix, and the frequency values in the Users-Locations matrix will be changed. The mentioned three processing steps and the users-similarity have to be proceeded as follows.

First, the similarity values between user i and j is computed using a cosine-based equation shown in Equation 1.

$$sim\left(\overrightarrow{i}, \overrightarrow{j}\right) = cos\left(\overrightarrow{i}, \overrightarrow{j}\right) = \frac{\overrightarrow{i} \cdot \overrightarrow{j}}{\| \overrightarrow{i} \|^2 * \| \overrightarrow{j} \|^2} \tag{1}$$

where $\cdot$ is the dot-product between the two user vectors.

The complexity in this step, when the re-calculation is to be proceeded, is shown as follows.

Determine user similarities: $O(m^2nr)$

where m is the number of users, n is the number of locations, and r is the number of activities.

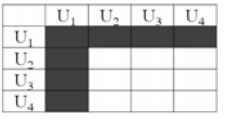


Fig. 3. The changing of rows and columns of the matrix when $user_1$ is updated

In this paper, we improve the efficiency of this step by re-calculating only the change. For example, when the data from $user_1$ is updated as shown in Figure 3. It can be seen that the update affects only the values related to $user_1$ as shaded area. The only update values of $\overrightarrow{i} \cdot \overrightarrow{j}$ and $\| \overrightarrow{i} \|$ related to changing users are computed and stored. Then, the values can be combined with the existing non-change values to form the new Users-Users matrix. Thus, the new complexity is shown as follows.

Determine user similarities: $O(m'm)$

where m' is the number of the users that has updated the frequency.

In the second step, the tensor decomposition is proceeded using PARAFAC decomposition equation as shown in Equation 2.

$$\begin{aligned} X &\leftarrow \tau_1(Z \odot Y)(Z^T Z * Y^T Y)^\dagger \\ Y &\leftarrow \tau_2(Z \odot X)(Z^T Z * X^T X)^\dagger \\ Z &\leftarrow \tau_3(Y \odot X)(Y^T Y * X^T X)^\dagger \end{aligned} \tag{2}$$

where τ_i is the mode-i tensor unfolding. $\odot$ is Khatri-Rao product. $*$ is entry-wise product. $\dagger$ is MoorePenrose pseudoinverse.

To calculate the decomposition, first, matrix Y and Z are generated to determine the matrix X. Then, the new matrix X and the matrix Z will be used to determine a new matrix Y. Last, the new matrix X and the new matrix Y will be used to determine a new matrix Z. The complexity of the algorithm is shown as follows.

Determine X: $O(mn^2r + n^3)$
Determine Y: $O(mn^2r + n^3)$
Determine Z: $O(mn^2r + n^3)$

In order to determine X efficiently, the existing $(Z \odot Y)(Z^TZ * Y^TY)^\dagger$ values can be used without any cost. Also, in the process of determining Y and Z, we can use the original value of Z^TZ, and the original value of X^TX respectively. Thus, we propose a new method of the tensor unfolding to update a location value in the matrices τ_1, τ_2, τ_3 as in Equation 3.

$$\begin{aligned} \tau_1[u, (a*n) - (n-l)] &= \tau_1[u, (a*n) - (n-l)] + 1 \\ \tau_2[l, (l*m) - (m-u)] &= \tau_2[l, (l*m) - (m-u)] + 1 \\ \tau_3[a, (a*m) - (m-u)] &= \tau_3[a, (a*m) - (m-u)] + 1 \end{aligned} \tag{3}$$

where u, l, a are mean the user u who do the activity a at the location l.

Obviously, such method update only the change, thus, it only costs $O(1)$ for each user-update. Thus, the new complexity of this step is shown as follows.

Determine X: $O(m'n^2r)$
Determine Y: $O(mn^2 + n^3)$
Determine Z: $O(mn^2r + n^3)$

Next, in the third step, the UCLAF algorithm is computed as shown in Equation 4.

$$\begin{aligned} \bigtriangledown_X L =& - A^{(1)}(Z \odot Y) + X[(Z^TZ) * (Y^TY)] + \lambda_1 L_B X + \lambda_4(XY^T - E)Y + \lambda_5 X \\ \bigtriangledown_Y L =& - A^{(2)}(Z \odot X) + Y[(Z^TZ) * (X^TX)] + \lambda_2(YU^T - C)U + \lambda_4(XY^T - E)^T X + \lambda_5 Y \\ \bigtriangledown_Z L =& - A^{(3)}(Y \odot X) + Z[(Y^TY) * (X^TX)] + \lambda_3 L_D Z + \lambda_5 Z \\ \bigtriangledown_U L =& \lambda_2(YU^T - C)^T X + \lambda_5 U \end{aligned} \tag{4}$$

where $A^{(i)}$ is the mode-i tensor unfolding. $\odot$ is Khatri-Rao product. $*$ is entry-wise product. B is the Users-Users matrix. C is the Locations-Features matrix. D is the Activities-Activities matrix. E is the Users-Locations matrix. $\lambda_1 - \lambda_5$ are the model parameters.

Its complexities are as follows.

Determine $\bigtriangledown_X L$: $O(mn^2r + m^2n + n^3)$
Determine $\bigtriangledown_Y L$: $O(mn^2r + n^2p + n^3)$

Determine $\bigtriangledown_Z L : O(mn^2r + n^3 + nr^2)$
Determine $\bigtriangledown_U L : O(n^2p)$

From Equation 4, it can be seen that the operations have very high dependency. Each change in the matrix X, Y, and Z from the step 2 changes every rows and columns in $\bigtriangledown_X L, \bigtriangledown_Y L, \bigtriangledown_Z L$, and $\bigtriangledown_U L$. So, only small computational cost can be reduced. That is, $A^{(1)}, A^{(2)}, A^{(3)}$ can be computed using Equation 3 in the same way as τ_1, τ_2, τ_3. Thus, the complexity in this part is decreased to $O(1)$ when a value is updated. Additionally, the matrix L_B, which is the Laplacian matrix of the Users-Users matrix, can be only updated for the changed user. However, the rest of the computation is needed to re-computed as the traditional approach. So, the new complexity for this step is shown as follows.

Determine $\bigtriangledown_X L : O(mn^2r + m^2n + n^3 + m'^2)$
Determine $\bigtriangledown_Y L : O(m' + mn^2r + n^2p + n^3)$
Determine $\bigtriangledown_Z L : O(m' + mn^2r + n^3 + nr^2)$
Determine $\bigtriangledown_U L : O(n^2p)$

In the final step, the tensor composition by the outer product equation is proceeded as shown in Equation 5. In which, we leave the computation intact as all the values can be changed from each user-update. The complexity of this step is $O(mnr + mn^2)$ which is exactly the same as the original UCLAF

$$[[X, Y, Z]] = \sum X \circ Y \circ Z \tag{5}$$

where $\circ$ denotes the outer product.

In conclusion, we improve the algorithm using the existing results from the previous round when it is possible. Thus, the same results as the re-calculation method can be obtained. Although, the proposed method can reduce the complexity, it requires the memory space for maintaining the previous results. However, the results from each step of calculation can be very sparse due to the nature of the recommendation. In this paper, we apply the sparse array [5] to store and maintain the results in order to utilize the space efficiently. The sparse array is an array that its data structure is developed to allocate memory only for the array indexes or positions which contain data or value. As a result, the sparse array can use less memory storage to implement our approach. The efficiency of the proposed approach as well as the effect of the sparse array will be evaluated in the next section.

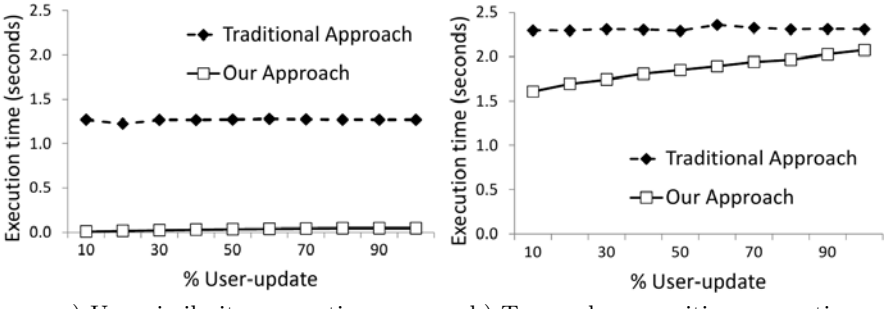
4 Experiments

In this section, we evaluate the computational and space efficiency of the proposed approach comparing with the efficiency of the re-calculation approach.

The dataset used in the experiments is obtained from the authors in [6], and it is prepared with the same method. The data consists of 164 users, 168 locations, 5 activities, and 13 features. In order to evaluate the results, we use

the dataset as the existing data repository, then the uniform-randomized user-update data are appended to the system. In which, the user-update data will be taken with the existing model to compute a new tensor. Then, the new tensor can be used for the recommendation. In each experiment, the percentage of the number of the user-updates is varied to evaluate its effect. The execution time of our approach covers the first model building until all the updates are finished. The experiments have been conducted on Intel Core i3 processor 4.4GHz with 4GB of main memory running Microsoft Windows 7. All the algorithms are implemented on Microsoft Visual C# Platform.

In the first experiment, we measure the efficiency of the user similarity computation in the first step of the model computing. The y-axis represents an execution time in second and the x-axis represents a percentage of user-updates. The result of the experiment is shown in Figure 4a). It is clear that the proposed algorithm is more efficient than the traditional approach. This is because the proposed algorithm can utilize the existing information which is stored to perform the calculation.

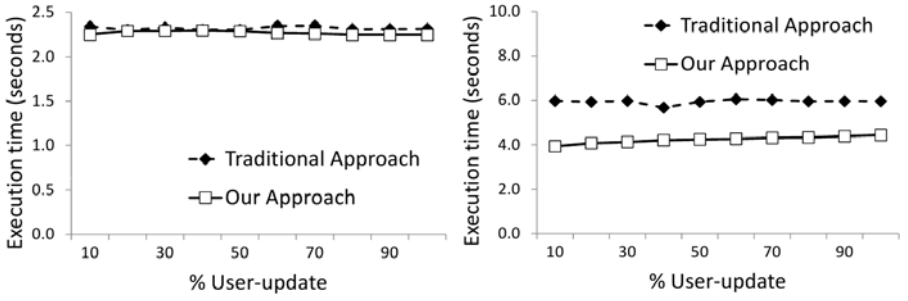


a) User similarity computing b) Tensor decomposition computing

Fig. 4. Results of the user similarity and tensor decomposition computation

In the second experiment, we measure the efficiency of the tensor decomposition computation in the second step. The result of the experiment is shown in Figure 4b). The efficiency of our proposed work is much more higher than the traditional re-calculation approach. The rationale behind this is that the complexity of τ_1, τ_2, τ_3 determination are only $O(1)$ for each user-update. Moreover, the proposed approach can utilize the existing values especially in the procedure of X determination, in which almost of the existing values can be re-used. However, the efficiency of our work degrades when the percentage of the user-update is increased. This is because the number of user-updates affects the computational cost directly, particularly the m' term.

Subsequently, the efficiency of the UCLAF algorithm is shown in Figure 5a). Evidently, the execution time of the proposed algorithm is rather similar to the re-calculation algorithm. As mentioned before, the high-dependency between each operation in this step does not allow us to improve the efficiency. Thus, the



a)UCLAF computing b) Overall computing

Fig. 5. Results of UCLAF algorithm computation and overall computation

efficiency gain is not much. Also, it can be seen that the amount of data to be updated from the users does not affect the execution time.

After the efficiency of our proposed work has been evaluated individually, the efficiency of the overall recommendation process is shown in Figure 5b). From the result, the proposed approach is obviously more efficient than the re-calculation approach. Although the execution time is slightly higher when the number of user-updates is increased due to the computation in second step. When combining all the calculation steps together for the recommendation, the efficiency degrade is reduced. Thus, it can be seen that our proposed work is highly efficient comparing with the traditional approach in term of the computational cost.

Last, the memory consumption of our approach is reported. The consumption between our proposed work with/without the sparse array technique, and the re-calculation approach are compared. The result is shown in Figure 6. Comparing with the traditional approach, our approach requires only 7% of the memory additionally. Furthermore, our approach with the sparse array technique requires only 3% of the additional memory. When considering the additional space cost with the efficiency gains from our approach, it can be concluded that our proposed work is very efficient.

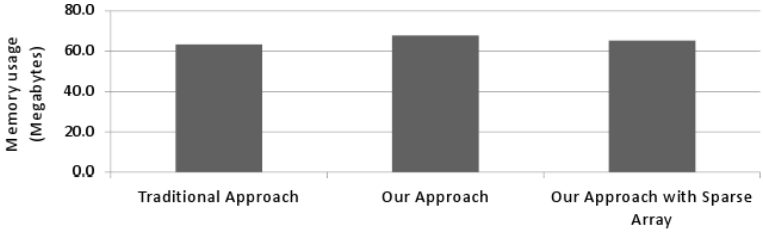


Fig. 6. Memory usage comparison

5 Conclusions and Future Work

In summary, we have proposed an efficiency improvement for the location-based recommendation system when the data are to be updated. The three steps of the recommendation, i.e. user similarity computation, tensor decomposition, and the UCLAF algorithm computation, are improved. The key idea is to utilize the existing computed data as much as possible. From the experiments, it can be seen that our proposed approach is much efficient than the traditional re-calculating approach. Also, our proposed with sparse array requires only small additional memory. Thus, the efficient approach can be implemented with not much space cost. In the future, we will further investigate the efficiency issues when the number of users, activities, locations, and features are changed. Such change can be occurred in real-life applications because some places may be closed or new popular locations and activities can be added. Addressing such issues can help improving the accuracy of the recommendation.

References

1. Domeniconi, C., Gunopulos, D.: Incremental support vector machine construction. In: Proceedings of the 2001 IEEE International Conference on Data Mining, pp. 589–592. IEEE Computer Society, Washington, DC, USA (2001)
2. Ganti, V., Gehrke, J., Ramakrishnan, R.: Demon–data evolution and monitoring. In: Proceedings of the 16th International Conference on Data Engineering (2000)
3. Resnick, P., Varian, H.R.: Recommender systems. Communication ACM 40, 56–58 (1997)
4. Sarwar, B., Karypis, G., Konstan, J., Reidl, J.: Item-based collaborative filtering recommendation algorithms. In: Proceedings of the 10th International Conference on World Wide Web, pp. 285–295. ACM, New York (2001)
5. Stoer, J., Bulirsch, R.: Introduction to numerical analysis. Texts in applied mathematics. Springer, Heidelberg (2002)
6. Zheng, V.W., Cao, B., Zheng, Y., Xie, X., Yang, Q.: Collaborative filtering meets mobile recommendation: A user-centered approach. In: Proceedings of the Twenty-Fourth AAAI Conference on Artificial Intelligence, AAAI 2010. AAAI Press (2010)
7. Zheng, V.W., Zheng, Y., Xie, X., Yang, Q.: Collaborative location and activity recommendations with gps history data. In: Proceedings of the 19th International Conference on World Wide Web, pp. 1029–1038. ACM, New York (2010)

Combining Topic Model and Co-author Network for KAKEN and DBLP Linking

Duy-Hoang Tran[1], Hideaki Takeda[2], Kei Kurakawa[2], and Minh-Triet Tran[1]

[1] Faculty of Information Technology,
University of Science Ho Chi Minh City, Vietnam
{tdhoang,tmtriet}@fit.hcmus.edu.vn
[2] National Institute of Informatics, Japan
{takeda,kurakawa}@nii.ac.jp

Abstract. The Web of Data is based on two simple ideas: to employ the RDF data model to public structured data on the Web and to set explicit RDF links to interlink data items within different data sources. In this paper, we describe our experience in building a system of link discovery between KAKEN, a database provides the latest information of research projects in Japan, and the DBLP Computer Science Bibliography. Using these links one can navigate from the information of a computer scientist in KAKEN to his publications in the DBLP database. Our problem of linkage between KAKE researchers and DBLP authors is name disambiguation. We proposed combining LDA based topic model and co-author network approach to improve linkage accuracy.

Keywords: Web of Data, LDA Model, Co-author Network, Connected Triple.

1 Introduction

The Web of Data is constantly growing over the last three years and has started to span data sources from a wide range of domains such as geographic information, people, companies, music, life-science data, books, and scientific publications. With the constant growth of scientific publications, bibliographic data sources become widespread. Linked datasets such as CiteSeer, ACM or DBLP are often consulted to find publications in a given domain or identify people working in an area of interest. KAKEN is database of Grants-in-Aid for Scientific Research contains the "Project Selected" documents and the research report summaries. The Grants-in-Aid for Scientific Research is granted whole field of science, and this database provides the latest information of the research projects in Japan exhaustively. KAKEN RDF allows asking sophisticated queries against datasets derived from KAKEN to other datasets on the Web. In this paper we will deal with the linkage problem between researchers in KAKEN and authors in DBLP.

Our challenge is the lack of associated properties of the entities in two data sources that should be link. We carry out analysis and evaluation of approaches which related to the data linkage problem and propose using SILK framework [4] to discover links between KAKEN researchers and DBLP authors based on their names. But the entity

J.-S. Pan, S.-M. Chen, N.T. Nguyen (Eds.): ACIIDS 2012, Part III, LNAI 7198, pp. 396–404, 2012.

names are often ambiguous. For example, the name "Hiroshi Suzuki" refers to 27 researchers in KAKEN. Also, the name "Hiroshi Suzuki" can be written "H. Suzuki" in DBLP. We list two types of ambiguity: (1) different researchers share the same name and (2) one author has different name aliases. We propose topic-based similarity measure and co-author network to improve the reliability of links. The main idea behind our solution is that if a researcher and an author are the same person, his papers and projects must be related to the same topic and they have some social relationships with the same person. We collect paper titles as a topic feature for an author and project titles as a topic feature for a researcher. We calculate the topic-based similarity of two features by LDA based topic model. Also, the researchers have co-member relationships with other member in the same project and the authors have co-author relationships with other author in the same paper. Using both co-member and co-author relationships we define a relationship-based network and use Connected Triple similarity to determine reliability of a link. The main results of this paper are (1) constructing a topic based similarity measure based on LDA model, (2) constructing a relationship based similarity measure based on co-author network approach, (3) building a system combining two measures to determine the reliability of a link, (4) linkage accuracy increase from 41,02% to 86,04%.

This paper is structured as follows: Section 2 briefly describes related work. Section 3 given an overview of data sources KAKEN and DBLP. Section 4 describes the system architecture with two major modules: LDA based similarity measure and Connected Triple similarity measure. Section 5 reports the results of implementation and review related work in Section 6.

2 Related Work

LinkedMDB [3] provides a demonstration of connecting several major existing movie web resources. Because the data sources are about movie, LinkedMDB chooses movie titles as feature to discover *owl:sameAs* links. The problem is matching only the titles may not be sufficient due to different representations of the same title. They use proper string similarity function and specific record matching techniques to achieve high accuracy. However, it is not easy to apply this idea to our problem because person name is more ambiguous than film title. SILK [4] use a declarative language for specifying which types of RDF links between data sources should be discovered as well as which conditions entities must fulfill in order to be linked. Depending on which data sources are linked, SILK has different thresholds ("accept" and "verify") for identifying similarity heuristics and qualifying the amounts of discovered links. This approach, however, only focuses on links of pairs of data sources: there is no guarantee that the information extracted from two data sources will be enough to find suitable entities in remains data sources. [5] present an algorithm to detect hidden *owl:sameAs* links or hidden relations in data sets. The main idea behind this solution is to extract useful features by applying supervised learning on frequent graphs. Then, using these extracted features to discover entities in data sources. This approach is not appropriate for our problem because it needs the existing links between data sources to discover the other hidden links.

3 Data Sources

KAKEN is the database which is established and provided by the National Institute of Informatics (NII) with support of the Ministry of Education, Culture, Sports, Science and Technology (MEXT) and the Japan Society for the Promotion of Science (JSPS). It is a part of "GeNii, the National Institute of Informatics academic content portal", established by NII and provides the latest information of the research projects in Japan. The database has more than 180,000 researchers and 2.7 millions projects in November 2010. DBLP (Digital Bibliography & Library Project) is a computer science bibliography website hosted at Universität Trier, in Germany. It was originally a database and logic programming bibliography site, and has existed at least since the 1980s. DBLP listed more than 1.3 million articles on computer science in January 2010. Journals tracked on this site include VLDB, a journal for very large databases, the IEEE Transactions and the ACM Transactions. Conference proceedings papers are also tracked. It is mirrored at five sites across the Internet. DBLP (L3S) is an effort to extract structured information from DBLP and to make this information available on the Web. DBLP (L3S) allows you to ask sophisticated queries against DBLP, and to link other data sets on the Web to DBLP data. This is a database published with D2R Server. It can be accessed using: (1) your plain old web browser, (2) Semantic Web browsers, (3) SPARQL clients.

4 System Architecture

The KAKEN and DBLP linking system have four main components. SILK framework is use to discover link between two data sources based on personal name. Then combining LDA based similarity and Connected Triple similarity to compute the similarity between two entities. Final, deciding a valid link if the similarity score above a threshold θ.

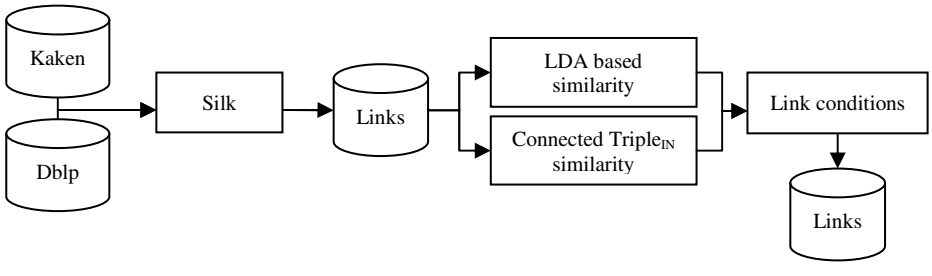


Fig. 1. KAKEN and DBLP linking system architecture

4.1 LDA Based Similarity

The main idea is that each author is associated with a multinomial distribution over topics and each topic is associated with a multinomial distribution over words. An

author with multiple documents is modeled as a distribution over topics. This is a generative model for document collections, the topic model, that simultaneously models the content of documents and the interests of authors. This generative model represents each document with a mixture of topics, as in state-of-the-art approaches like Latent Dirichlet Allocation (Blei et al., 2003), and extends these approaches to author modeling by allowing the mixture weights for different topics to be determined by the authors of the document. By learning the parameters of the model, we obtain the set of topics that appear in a corpus and their relevance to different documents, as well as identifying which topics are used by which authors. The algorithm has three steps:

Step 1: finding hidden topic by using Latent Dirichlet Allocation, training data is list of document, each document is list of paper title in a conference.
Step 2: extract the entity topic feature: an author is featured as the list of paper titles; a researcher is featured as the list of project titles. Appling LDA topic model for these features we have vectors that each dimension corresponds to probability that an author or researcher is related to a hidden topic.
Step 3: using cosine similarity to compare similarity between two vector, from which determine that the researcher and the author is the same person.

4.2 Connected Triple$_{IN}$ Similarity

Assume that all researcher and author that have the same name is a same person. The researchers that are member of the same project will have co-member relationships. The author will have the co-author relationships with other author that have the same paper. Using both co-member and co-author relationships we have a network G in which authors/researchers are represented as vertices V , and relationships builds the edges E. Then for each link researcher and author, we calculate the Connected Triples Similarity of two entities. A Connected Triple $\wedge = \{V_\wedge, E_\wedge\}$ can be described as a sub graph of G consisting of three vertices with:

$$V_\wedge = \{A_1, A_2, A_3\} \subset V \text{ and } E_\wedge = \{e_{A1,A2}, e_{A1,A3}\} \in E, \{e_{A2,A3}\} \notin E.$$

The edges in the co-author network will be weighted according to Liu et al. [3]. With $V = \{v_1,...,v_n\}$ as the set of n authors, m the amount of publications $A = \{a_1,...,a_k,...a_m\}$ and f (a_k) the amount of authors of publications a_k the weight between two authors v_i and v_j for publications a_k is calculated by:

$$g(i,j,k) = \frac{1}{f(a_k)-1} \quad (1)$$

There by the weight between two authors for one publication is smaller the more authors collaborated on this publication. Considering the amount of publications two authors i and j collaborated on together, an edge between these authors is calculated with (2) which leads to higher weights the more publications the two authors share.

$$C_{ij} = \sum_{k=1}^{m} g(i,j,k) \quad (2)$$

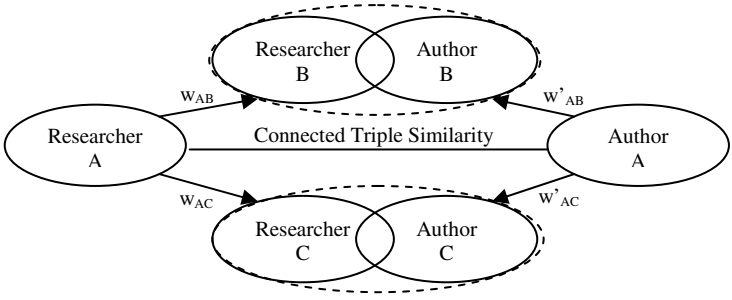


Fig. 2. Combining co-member and co-author network

Applying a normalization the weight between two authors i and j considering the amount of co-authors and publications is calculated by (3) leading to a directed co-author graph.

$$w_{ij} = \frac{C_{ij}}{\sum_{r=1}^{n} C_{ir}} \tag{3}$$

The similarity of two authors using Connected Triples can consequently be either calculated on incoming edges or outgoing edges:

$$ConnectedTriple_{in} = \sum_{\forall c \in V \text{ with } e_{ci}, e_{cj} \in E, e_{ij} \notin E} w_{ci} + w_{cj} \tag{4}$$

$$ConnectedTriple_{out} = \sum_{\forall c \in V \text{ with } e_{ic}, e_{jc} \in E, e_{ij} \notin E} w_{ic} + w_{jc} \tag{5}$$

5 Experimental Result

5.1 LDA Training Data

The LDA based topic model data training is a list of documents, each document is a list of paper titles in a conference that extract from DBLP. In this problem, we use 10.245 conferences as the training data. The estimation of number of hidden topic k is very important. If k is too large each topic feature will be narrow, if k is too small each topic feature will be wide. We experiment the parameters k with 50 topics, 100 topics and 150 topics. The testing is performed on the data set includes 724 links that have been labeled by manual. In our experiment, we randomly divided test data into three sets to determine whether the parameters are consistent with local data. We set the threshold θ for the measure, if the similarity score is greater than the threshold we will conclude that the two entities will be linked. We need to inspect different thresholds to find the optimal threshold.

Fig. 3 show the accuracy obtained with different value of number of hidden topic and threshold. Note that accuracy reported is percentage of both true positive and negative. Base on the result we chose number of hidden topic k=100 topics and threshold θ=0.7. Fig.4 shows that the threshold is consistent with local data.

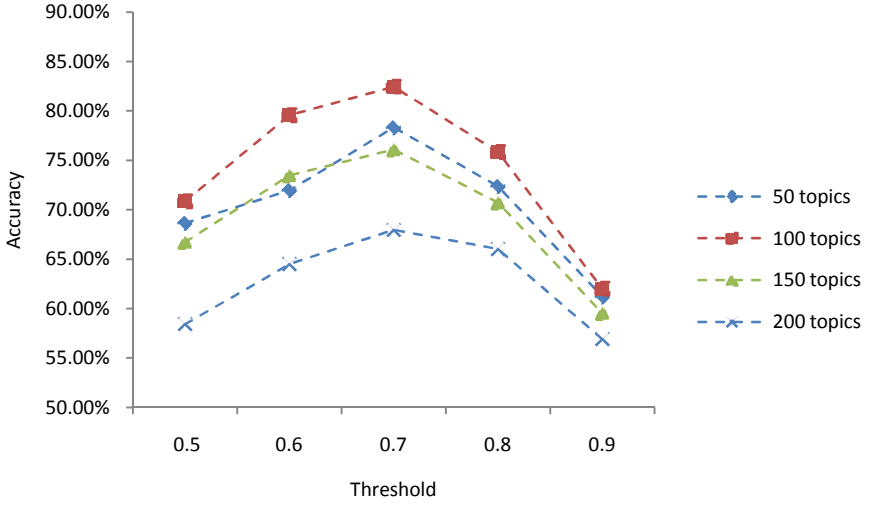


Fig. 3. LDA-based similarity accuracy

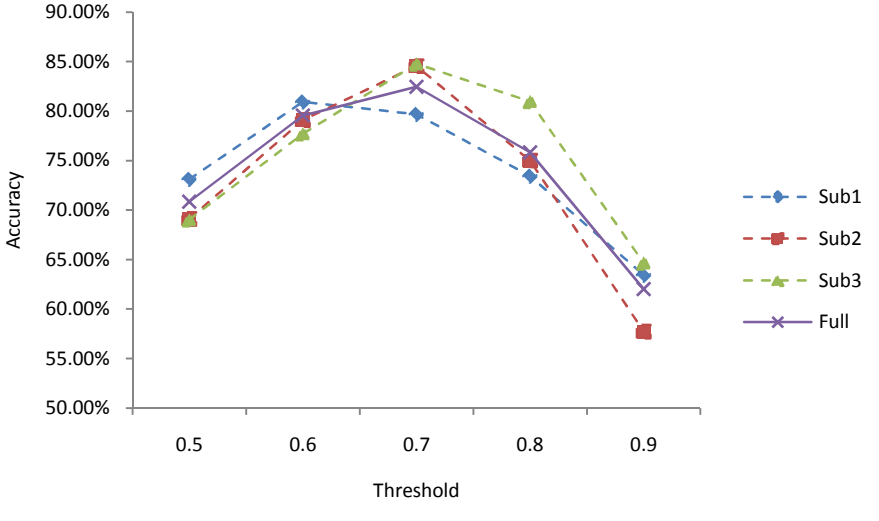


Fig. 4. LDA-based similarity accuracy in sub datasets

5.2 Compare LDA Model to pLSI and TFIDF Model

Fig. 5 compares the accuracies of LDA topic model and pLSA topic model also TFIDF weight and shows that LDA gave a better result than the others.

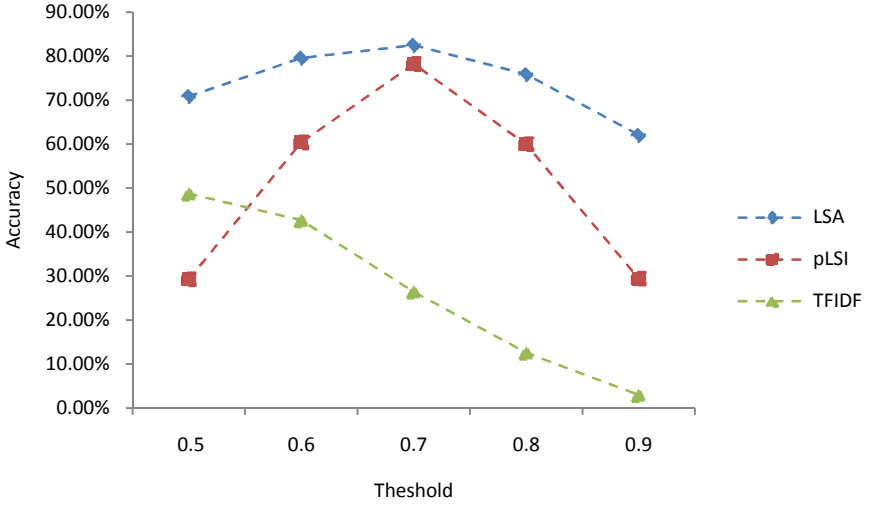


Fig. 5. Accuracies of LDA and pLSA, TFIDF

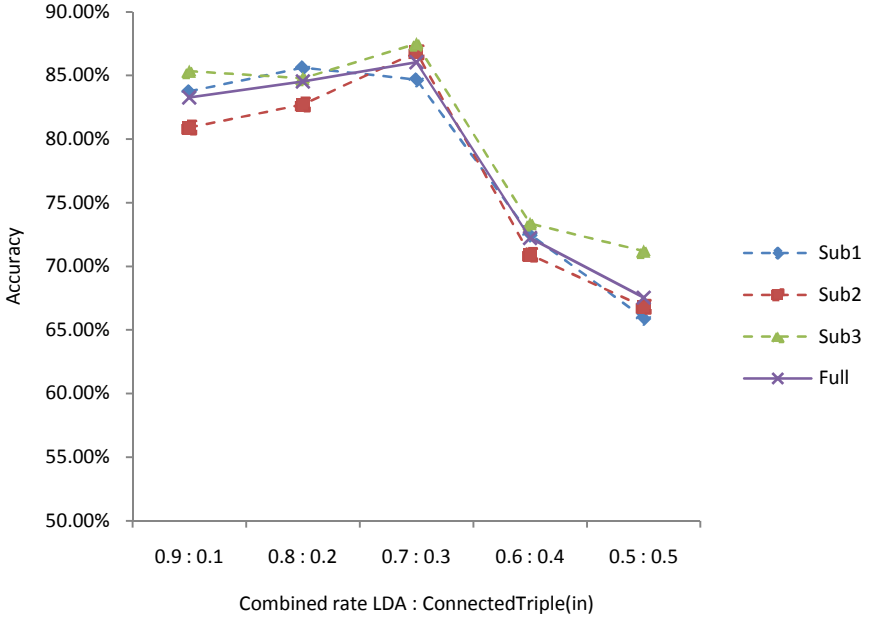


Fig. 6. Combining LDA topic similarity and ConnectedTriple similarity

5.3 Combining Two Approaches

Fig. 6 show the accuracy obtained with different combined rate between LDA topic similarity and Connected Triple similarity in different sub datasets. We find the optimal rate is 0.7:0.3. Base on these results we chose the threshold θ=0.7 and the rate of combining two similarity measures is 0.7:0.3. Fig. 7 shows that the combining of LDA base similarity and Connect Triple similarity gave a better result.

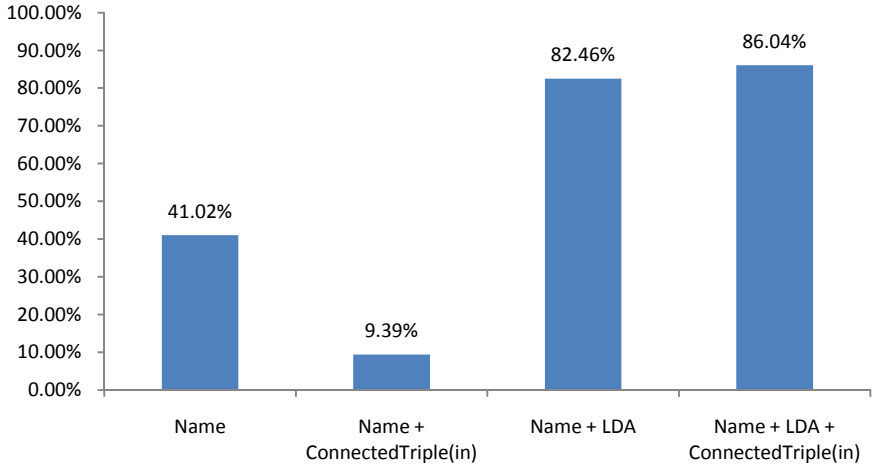


Fig. 7. Accuracy of combining similarity measures

6 Conclusions

We presented an approach to solve the name ambiguous problem in KAKEN and DBLP linkage. Our solution is combining LDA based topic model and co-author network approach to improve accuracy of the links. We compared the LDA-based topic model with other models including pLSA and TFIDF weight. The paper has contributed a small part in solving the bibliographic data linkage and applying for a specific problem. The results increase the accuracy from 41.02% to 86.04%.

References

1. Bizer, C., Heath, T., Idehen, K., Berners-Lee, T.: Linked data on the web. In: Proceeding of the 17th International Conference on World Wide Web, WWW 2008 (2008)
2. Bizer, C., Heath, T., Ayers, D., Raimond, Y.: Interlinking Open Data on the Web. In: Demonstrations Track, 4th European Semantic Web Conference, Innsbruck, Austria (2007)
3. Hassanzaded, O., Consens, M.: Linked movie data base. In: Proceedings of the WWW 2009 Workshop on Linked Data on the Web, Madrid, Spain (2009)

4. Volz, J., Bizer, C., Gaedke, M., Kobilarov, G.: Silk - a link discovery framework for the web of data. In: Proceedings of WWW 2009 Workshop on Linked Data on the Web, Madrid, Spain (2009)
5. Le, N.T., Ichise, R., Le, H.B.: Detecting Hidden Relations in Geographic Data. In: Proceedings of the 4th International Conference on Advances in Semantic Processing, Florence, Italy (2010)
6. Biryukov, M.: Co-Author Network Analysis in DBLP: Classifying Personal Names. In: 2nd International Conference on Modeling, Computation and Optimization in Information Systems and Management Sciences, Metz, France (2008)
7. Rosen-Zvi, M., Griffit, T., Steyvers, M., Smyth, P.: The Author-Topic Model for Authors and Documents. In: 20th Conference on Uncertainty in Artificial Intelligence, Banff, Canada (2004)
8. Reuther, P., Walter, B., Ley, M., Weber, A., Klink, S.: Managing the Quality of Person Names in DBLP. In: Gonzalo, J., Thanos, C., Verdejo, M.F., Carrasco, R.C. (eds.) ECDL 2006. LNCS, vol. 4172, pp. 508–511. Springer, Heidelberg (2006)
9. Blei, D., Ng, A., Jordan, M.: Latent Dirichlet Allocation. Journal of Machine Learning Research (JMLR) 3, 993–1022 (2003)

PLR: A Benchmark for Probabilistic Data Stream Management Systems

Armita Karachi, Mohammad G. Dezfuli, and Mostafa S. Haghjoo

Computer Engineering Department,
Iran University of Science and Technology,
Tehran, Iran
armita_karachi@comp.iust.ac.ir,
{mghalambor,haghjoom}@iust.ac.ir

Abstract. Inherent imprecision of data streams in many applications leads to need for real-time uncertainty management. The new emerging Probabilistic Data Stream Management Systems (PDSMSs) are being developed to handle uncertainties of data streams in real-time. Many approaches have been proposed so far but there is no way to compare them regarding precision and efficiency. This problem motivated us to design an evaluation framework to compare performance and accuracy of PDSMSs with each other and also with probabilistic databases. In this paper, after a brief introduction to PDSMSs, we describe requirements and challenges for designing a PDSMS benchmark. Then, we present different parts of our framework including probabilistic data stream generator, queries, and result evaluator. Furthermore, we focus on implementation aspects and use our framework to evaluate effects of floating precision in our PDSMS prototype.

Keywords: PDB, PDSMS, Uncertain Stream, Benchmark.

1 Introduction

In many applications, data streams are inherently uncertain. Some examples are monitoring sensor networks [1], RFID networks [2, 3], GPS systems [4], camera sensor networks [5], radar sensor networks [6], and scientific data. For example, measured values in monitoring sensor networks have errors. Almost all scientific data including model outputs, estimates, experimental measurements, and hypothetical data is fundamentally uncertain [7]. Handling uncertainties is potentially complex and inefficient for end users. Due to significant need for supporting uncertain data, many probabilistic databases have been developed such as Trio [8], Orion [9], MayBMS [10], and MystiQ [11]. However, these systems cannot perform on continuous probabilistic data streams in a real-time fashion. There is also some work on managing probabilistic data streams, but it is mostly about algorithms instead of a complete system (e.g. CLARO [12, 13]). We are working on filling this gap by developing the first Probabilistic Data Stream Management System (PDSMS) as an extension of our existing DSMS [14].

J.-S. Pan, S.-M. Chen, N.T. Nguyen (Eds.): ACIIDS 2012, Part III, LNAI 7198, pp. 405–415, 2012.

One of the real problems toward developing PDSMSs is lack of a benchmark to evaluate their algorithms and to compare them with each other. PDSMSs are still state of the art systems and inconsistent. Therefore, it is too early to present a proper benchmark for them. However, we still need a framework to do our evaluations and develop first prototypes. We extended the existing *linear road* benchmark [15] of DSMSs and made it probabilistic to have an evaluation framework for new emerging PDSMSs. We call it PLR (Probabilistic Linear Road). Three main parts of PLR are: 1) probabilistic input data stream generation, 2) queries, and 3) query results evaluation. We describe each part in more details. As part of our ongoing work, we are extending PLR to address correlations and more distributions.

The rest of the paper is organized as follows. In Section 2 we focus on major requirements and background of benchmarking PDSMSs as well as existing related work. In Section 4 we discuss about input data streams in our evaluating framework. In Section 5, we define some queries for PLR. In Section 6 we introduce evaluation metrics of PLR. Section 7 is about empirical study of PLR. Finally we conclude this paper in Section 8 and propose our future work.

2 Background and Related Work

Two main categories for uncertainty management are: 1) fuzzy-based modeling [16], and 2) probabilistic-based modeling [7, 12, 13, 17]. Since nature of data in many PDSMS applications is probabilistic [18], we focus on probabilistic-based approach. To model probabilistic values in probability-based models, we can use discrete or continuous distributions. Discrete models can capture continuous uncertainty by sampling and discretization. Most of the work on probability-based models only covers discrete distributions [12]. Supporting continuous distributions in a system poses more complexity and overhead in query processing [12]. However, there exists a few systems which support continuous distributions natively [9, 13]. Note that in discrete models, there is a trade-off between accuracy and efficiency. Uncertainty also can be at tuple or attribute level. In tuple uncertainty [11, 19, 20], presence of a tuple in a stream is probabilistic. On the other hand, attribute uncertainty [21] covers probable values for each probabilistic attribute. In PLR, we cover both tuple-level and attribute-level uncertainties.

In addition to data model and uncertainty level, there are different query processing approaches divided into two main categories: 1) computing the whole probability density function (pdf) of the results, and 2) using some statistical information. As stated in [12], using pdf is more reliable and efficient in most of the applications in order to gain useful information from intermediate results. In PLR, we support both discrete pmfs and continuous pdfs. We use Categorical discrete distributions for discrete attributes and Gaussian distributions for continuous attributes. These two distributions are the most common distributions supported by all work done so far [4, 7, 8, 9, 12, 13].

Designing a benchmark for a PDSMS has many noticeable challenges due to streaming and probabilistic characteristics of data. Four main reasons for difficulty of PDSMS benchmarking are: 1) inconsistency between PDSMS models, 2) continuous queries, 3) many correct results, and 4) lack of a standard query language.

There are numerous domain-specific database benchmarks [22], e.g., for OLTP (TPC-C), decision support (TPC-H, TPC-R, APB-1), information retrieval, spatial data management (Sequoia), stream-based data management [15], and some others such as [23, 24, 25, 26, 27]. However, only two benchmarks exist for DSMSs. The first one is *NEXMark* [28] (Niagara Extension to *XMark* [29]) which is about a simple online auction. The second one is *linear road* [15] which was the base of our evaluation framework. In spite of all existing benchmarks, there is no benchmark for probabilistic databases because these systems are still very inconsistent and state of the art [30].

3 Probabilistic Input Data Streams in PLR

Input data streams in PLR is are probabilistic versions of data streams in *linear road*. Input tuples are from three types of: 1) position reports, 2) toll requests, and 3) account balance requests. Position reports are primary tuples issued every 30 seconds by vehicles specifying their current *speed* and *position* on the expressways (Fig. 1). Toll requests are ad-hoc one-time queries received from traffic monitoring systems to find last tolls assigned to different parts of the expressways. Account balances are ad-hoc one-time query requests issued by drivers to get their current account balances. Each time a vehicle issues a position report, it may also send an ad-hoc one-time query with probability of one percent. The three mentioned input streams are multiplexed together as a single stream. Fig. 1 illustrates a segment in PLR expressways.

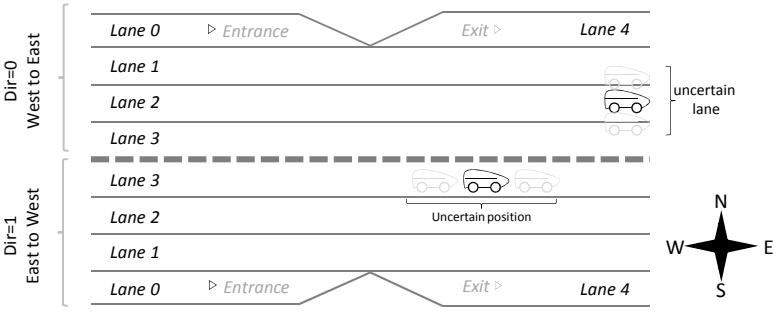


Fig. 1. A segment in PLR expressways

3.1 Input Schema

Input data streams in *linear road* are certain; therefore, we have designed a convertor to make the input data probabilistic. In generating uncertainty we should focus on two issues: 1) domains of uncertainty, and 2) the form of uncertainty used in each domain. In PLR, two main domains of uncertainties are vehicles' *position* and *speed* which are not accurate due to many factors such as sensors malfunction and environmental conditions. Among different fields of location, in horizontal coordinate, position is not accurate and in vertical coordinate, lane number is uncertain. Position and speed cannot be represented by discrete distributions. Thus, we used Gaussian distribution

for position and speed with average value equal to reported value and a variance according to supposed sensors' error bounds. For each position tuple, there are four possibilities for the vehicle's lane number. Therefore, we use Categorical discrete distribution for representing lane uncertainty. For tuple-level uncertainty we define an attribute which indicates the tuple existence probability. The input schema for attributes of position report in PLR is illustrated in Table 1.

Table 1. Attributes of position report in PLR

Input Fields	Description
Type	identifies the tuple as (0:Position Report, 1:Toll Request, and 2:Account Balance Request)
Time	Seconds since start of simulation
VID	vehicle identifier (0 . . . 999,999)
Spd	vehicle speed (0 . . . 100 MPH), Gaussian: G(mean, var)
XWay	Expressway number (0 . . . 9) = [y/17600]
Lane	lane of the expressway (0/4:entrance/exit ramp, 1 – 3:travel lane) = [(17599−(y mod 17600))/8], Categorical distribution: (Lane0, Probability0) | ...| (Lane3, Probability3)
Dir	Direction of the expressway (0:Westbound, 1:Eastbound) = [Lane/4]
Pos	horizontal position of the vehicle (0 . . . 527999 feet), Gaussian distribution: G(mean, var)
Conf	Tuple existence probability (0,1)

4 Probabilistic Queries in PLR

PLR has a set of well-defined queries on its probabilistic data streams to determine overall performance of PDSMSs. The main focus is on primary operations such as select, project, join and aggregation. We cover joining input data with various types of uncertainties such as Gaussian-Gaussian, Gaussian-Categorical and Categorical-Categorical distributions. SUM and AVG operators are used for probabilistic attributes. We support both predefined continuous and ad-hoc one-time queries. We used CQL [31] to specify our queries. Each of these queries must be answered accurately and within before a response time deadline.

4.1 Toll Query

Toll query is about computing tolls and maintaining toll accounts every time a vehicle reports a position in a new segment. Every time a position report shows a vehicle entering a new segment, the toll reported for that segment will be charged to the vehicle's account. As the goal of variable tolling is to prevent congestion in different parts of expressways, toll computation for a segment is based on three factors: 1) number of vehicles, 2) average speed, and 3) nearest upstream accident. We should first determine how to calculate these three factors and then how they affect toll computation.

As discussed before, *position* attribute is not a certain field. Thus, to find the number of vehicles in each segment, probability of each vehicle being in that segment should be computed. An Aggregation operator is required to calculate sum of these probabilities as the number of vehicles in that segment. Query 1 computes the number of vehicles in each segment.

Query 1. Number of vehicles in a segment.

```
SELECT      Seg#, Count(*) as NV
FROM        Pos [PARTITION BY VID ROWS 1]
GROUP BY    SegNum(XWay, Dir, Pos) As Seg#
```

In the above query, *SegNum(.)* is a simple function which generates a unique number to specify a single segment. The other attribute which is not certain is *speed*. Each position report has a speed represented by Gaussian distribution. To calculate the average speed in each segment, an AVG operator is needed on probabilistic speed attributes. In addition, A Join operator is also required to join the results of aggregation operators based on segment number. Query 2 computes the average speed in each segment.

Query 2. Average speed for each segment

```
SELECT      Seg#, Avg(Spd) as Spd
FROM        Pos [PARTITION BY VID ROWS 1]
GROUP BY    SegNum(XWay, Dir, Pos) As Seg#
```

After finding number of vehicles and average speed, we should determine how these parameters affect toll computation. We use Eq. 1 to compute a basic toll value, T_{base}, based on number of vehicles (NV) and average speed (Spd). We use this value in addition to accident information to compute final toll value in Eq. 2. Query 3 computes toll based on intermediate results of Query 1,2.

$$T_{base} = \alpha \log_k \frac{NV + 1}{Spd + 1} \tag{1}$$

Query 3. Toll computation

```
SELECT   *, TollClac(Spd.Spd, Num.NV, SegDistance(Sg#, A1.Seg#))
FROM     (SELECT
          Num.Seg# as Sg#, Num.NV, Spd.Seg# as Sg#, Spd.Spd
           FROM    Num,Spd
           WHERE Num.Seg# = Spd.Seg#)
LEFT OUTER JOIN Accident A1
ON       Sg# <= A1.Seg#
AND      SegDistance(Sg#, A1.Seg#) < 5
AND      SameXwayDIR(Sg#, A1.Seg#)
WHERE NOT EXIST (SELECT *
                         FROM Accident A2
                         WHERE
                                  SegBetween(A2.Seg#, Sg#, A1.Seg#))
```

There are five functions in toll query. *TollClac(.)* is a function used for toll calculation described in Eq. 2 (H_{Acc} will be discussed in 4.2). This function calls another function, *SegDistance(.)*, to detect upstream accidents which affect toll processing. Note that only accidents within 5 upstreams are important in toll calculation; Therefore, *SegDistance(.)* finds the distance between two segments. When accidents are detected, we use *SameXwayDIR(.)* function to check equality of expressway numbers and directions. *SegBetween(.)* function checks if there is any closer accident.

$$Toll = \alpha_1 T_{base} + \alpha_2 H_{Acc} + C \quad (2)$$

The other parameter in toll calculation is accidents which will be described in the next subsection. α parameters and C are only for biasing formulas.

4.2 Accident Query

The other continuous query is detecting and reporting accidents. An accident is detected whenever two or more vehicles are stopped in the same location. A vehicle is considered stopped whenever two consecutive position reports of that vehicle indicate the same location. Query 4 represents the CQL version of this query.

Query 4. Accident

```
SELECT  s1.Seg,Pos, Max(s1.Time,s2.Time)
FROM    Stop s1, Stop s2
WHERE   s1.Seg = s2.Seg
AND     s1.Pos = s2.Pos
AND     s1.VID <> s2.VID
AND     Abs(s1.Time - s2.Time) < 60
```

In accident query the first step is to identify stopped vehicles described in Query 5.

Query 5. Stopped vehicles

```
CREATE STREAM stop AS
SELECT
        p1.VID, p1.Seg, p1.Pos, Max(p1.Time,p2.Time) as Time
FROM    pos p1, pos p2
WHERE   p1.VID = p2.VID
AND     p1.Seg = p2.Seg
AND     p1.Pos = p2.Pos
AND     p1.Time <> p2.Time
AND     Abs(p1.Time - p2.Time) < 60
```

Accident detection also affects toll computation as mentioned before. Eq. 3 computes a complementary parameter for toll value with regard to distance between current segment and detected accident in front. α and x are biasing parameters and d is the distance to the next detected accident.

$$H_{Acc} = \alpha x^{-d} \quad (3)$$

5 Evaluation Metrics

Evaluating probabilistic data stream processing systems is more challenging and complex than PDBs or DSMSs. The most important metrics are: 1) response time and 2) accuracy of results. We use these two metrics because we expect PDSMSs to produce real-time accurate results. In PLR we determine a specific grade based on response time and accuracy parameters. Systems achieving higher grade (less query response time and at the same time greater grade in accuracy) are more appropriate in processing probabilistic data streams.

5.1 Response Time

One of the important goals of PDSMSs is time efficiency. PDSMSs should insert in and out system timestamps in tuples. After computing the elapsed time between these two timestamps, the evaluator can report response time metrics such as min, max, and average response time and variance, as well as, the percentage of results which missed the specified deadlines. Similar to *linear road*, we declare a 30 sec. response time deadline for toll query and 15 sec. for accident query.

5.2 Accuracy

In PDSMSs, input data streams and output results are probabilistic so that accuracy is very important. To evaluate accuracy, we need the correct results. We cover both tuple-level and attribute-level uncertainties therefore, accuracy should be evaluated according to these two uncertainty levels. For attribute-level uncertainty, the precision of each probabilistic attribute should be measured separately. Accuracy of the whole tuple is based on these measurements. We find the difference between reported and ideal attribute values to measure mean and variance of deviation. Note that attribute values are distributions rather than simple values and therefore we used the notion of variation distance between distributions [32].

6 Empirical Study

In this section, we present an empirical study of PLR. The implementation is on our PDSMS prototype [14]. At first, we intended to make a comparison between a PDSMS and a PDB. The only compatible PDB for PLR is Orion [9] which supports continuous distributions. Unfortunately, current version of Orion (ver. 2) does not support aggregation and join on continuous probabilistic attributes. Therefore, we postpone this comparison for the extended version of this paper and only present a simple example of implementing PLR. We show how changing precision affects response time and accuracy in our PDSMS. By this way, we can choose a proper precision for system tuning.

6.1 Implementing PLR on Sarcheshmeh

Our input data streams contains about 130 ,000 position reports for 15 minutes. All of these position reports are for a single expressway. The input rate is also non-decreasing during this time.

One of the most important parameters in PDSMSs is precision. We define precision here as a probability threshold which probability values less than it would be considered 0 (i.e. impossible). Each application can set its precision. Here we show how changing precision will affect different QoR and performance metrics. It is also a motivation for overload controlling (adaptation and load shedding) in PDSMSs using floating precision. Figures 2 and 3 illustrate the effect of precision on toll error. Figure 2 shows the effect of most important precision candidates on toll error during 15 minutes experiment. As you see in the chart, it increases over time slowly and it seems that $p = 0.01$ (precision equal to 0.01) is as good as $p = 0$. Figure 3 gives us a better understanding about the effect of precision on mean toll errors. As you see in the chart, mean toll error increases drastically by decreasing precision (i.e. increasing p value). The effect of changing precision on periodic response time, as a performance metric, illustrated in figures 4 and 5. More instability for $p = 0.5$ is because of the less number of tuples. If we consider two mentioned metrics together, $p = 0.01$ is a good candidate for precision as it leads to good results for both toll error and response time.

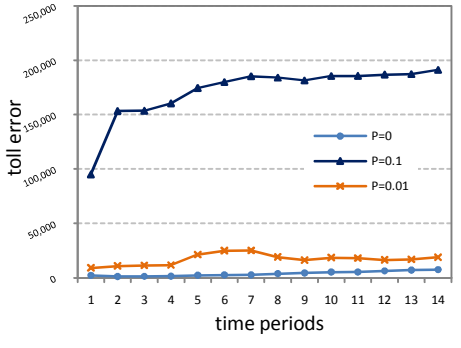


Fig. 2. The effect of changing precision on periodic toll errors

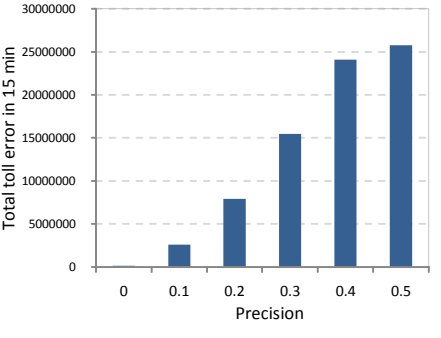


Fig. 3. The effect of changing precision on mean toll errors

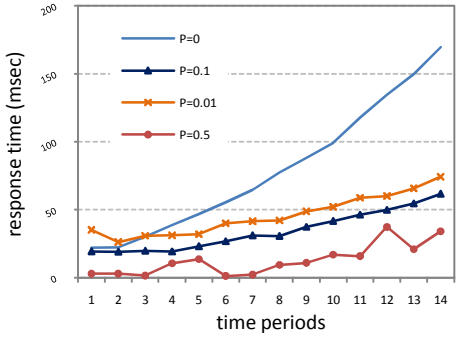


Fig. 4. The effect of changing precision on periodic response time

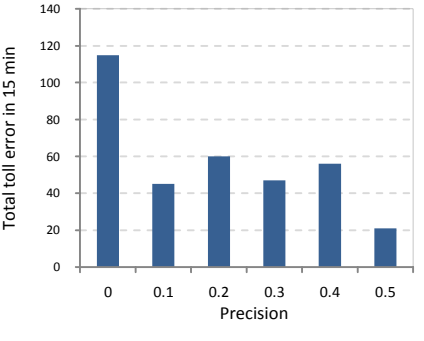


Fig. 5. The effect of changing precision on mean response time

7 Conclusion and Future Work

There is a growing attitude toward real-time querying of probabilistic data streams. However, no benchmark is available for comparing the performance of PDSMSs relative to each other and also to PDBs. In this paper, after a brief introduction to probabilistic data stream management, we presented our evaluations framework, PLR, based on *linear road* benchmark. We made an extension instead of a new benchmark because it is too early to introduce a new benchmark for PDBs and PDSMSs. In *linear road*, input data is accurate without any uncertainty which is unrealistic in practice. We covered data uncertainty as a first-class concept in PLR and provided a convertor to make input data streams probabilistic. Attribute and tuple-level uncertainties are included in PLR by using both discrete and continuous distributions. PLR includes both continuous and one-time queries for detecting accidents, toll calculation, penalty computation, and also toll alerts. We also developed an evaluation tool to evaluate PDSMS results based on PLR metrics.

PLR meets general benchmark requirements of relevance, portability, scalability and simplicity as well as the probabilistic requirements discussed in section 2. We also implemented PLR on our PDSMS prototype to check its effectiveness in action. PLR also helps us to better develop *Sarcheshmeh*. We invite others to run PLR on their own systems. We leave supporting correlations and more distributions for our future work. We also intend to promote PLR to a PDSMS benchmark in near future.

Acknowledgments. This work was supported by Iran Research Institute for ICT (ITRC) under grant No. 500/5366.

References

1. Deshpande, A., Guestrin, C., Madden, S., Hellerstein, J.M., Hong, W.: Model-Driven Data Acquisition in Sensor Networks. In: VLDB, pp. 588–599 (2004)
2. Jeffery, S.R., Franklin, M.J., Garofalakis, M.N.: An Adaptive RFID Middleware for Supporting Metaphysical Data Independence. VLDB Journal 17(2), 265–289 (2007)
3. Welbourne, E., Khoussainova, N., Letchner, J., Li, Y., Balazinska, M., Borriello, G., Suciu, D.: Cascadia: A System for Specifying, Detecting, and Managing RFID Events. In: MobiSys, pp. 281–294 (2008)
4. Kanagal, B., Deshpande, A.: Online Filtering, Smoothing and Probabilistic Modeling of Streaming Data. In: ICDE (2008)
5. Kulkarni, P., Shenoy, P., Ganesan, D.: Approximate Initialization of Camera Sensor Networks. In: Langendoen, K.G., Voigt, T. (eds.) EWSN 2007. LNCS, vol. 4373, pp. 67–82. Springer, Heidelberg (2007)
6. Kurose, J., Lyons, E., McLaughlin, D., Pepyne, D., Philips, B., Westbrook, D.L., Zink, M.: An End-User-Responsive Sensor Network Architecture for Hazardous Weather Detection, Prediction and Response. In: Cho, K., Jacquet, P. (eds.) AINTEC 2006. LNCS, vol. 4311, pp. 1–15. Springer, Heidelberg (2006)

7. Singh, S., Mayfield, C., Shah, R., Prabhakar, S., Hambrusch, S., Neville, J., Cheng, R.: Database Support for Probabilistic Attributes and Tuples. In: Proc. of the IEEE 24th International Conference on Data Engineering (2008)
8. Agrawal, P., Widom, J.: Continuous Uncertainty in Trio. In: MUD (2009)
9. Singh, S., Mayfield, C., Mittal, S., Prabhakar, S., Hambrusch, S., Shah, R.: Orion 2.0: Native Support for Uncertain Data. In: Proc. of ACM SIGMOD (2009)
10. Huang, J., Antova, L., Koch, C., Olteanu, D.: MayBMS: a Probabilistic Database Management System. In: Proc. of the 35th SIGMOD (2009)
11. Re, C., Suciu, D.: Managing Probabilistic Data with MystiQ: The Can-Do, the Could-Do, and the Can't-Do. In: Greco, S., Lukasiewicz, T. (eds.) SUM 2008. LNCS (LNAI), vol. 5291, pp. 5–18. Springer, Heidelberg (2008)
12. Diao, Y., Li, B., Liu, A., Peng, L., Sutton, C., Tran, T., Zink, M.: Capturing Data Uncertainty in High-Volume Stream Processing. In: CIDR (2009)
13. Tran, T.T., Peng, L., Li, B., Diao, Y., Liu, A.: PODS: a New Model and Processing Algorithms for Uncertain Data Streams. In: Proc. of the International Conference on Management of Data, Indianapolis, Indiana (2010)
14. Haghjoo, M.S., Dezfuli, M.G., Azizjalali, A.: Designing a Probabilistic Data Stream Management System. International Review on Computers and Software 5(6) (2010)
15. Arasu, A., Cherniack, M., Galvez, E.F., Maier, D., Maskey, A., Ryvkina, E., Stonebraker, M., Tibbetts, R.: Linear Road: a Stream Data Management Benchmark. In: VLDB Conference, Toronto (2004)
16. Galindo, J., Urrutia, A., Piattini, M.: Fuzzy Databases: Modeling, Design, and Implementation. Idea Group Publishing (2006)
17. Liu, H., Hwang, S., Srivastava, J.: Probabilistic Stream Relational Algebra: a Data Model for Sensor Data Streams. Technical Report, University of Minnesota (2004)
18. Faradjian, A., Gehrke, J., Bonnet, P.: GADT: a Probability Space ADT for Representing and Querying the Physical World. In: ICDE (2002)
19. Benjelloun, O., Sarma, A.D., Halevy, A., Widom, J.: ULDBs: Databases with Uncertainty and Lineage. In: Proc. of the 32nd International Conference on VLDB, pp. 953–964 (2006)
20. Deshpande, A., Madden, S.: MauveDB: Supporting Model-based User Views in Database Systems. In: Proc. of ACM SIGMOD, pp. 73–84 (2006)
21. Barbará, D., Garcia-Molina, H., Porter, D.: The Management of Probabilistic Data. IEEE Transactions on Knowledge and Data Engineering 4(5), 487–502 (1992)
22. Böhme, T., Rahm, E.: XMach-1: a Benchmark for XML Data Management. In: Datenbanksysteme in BÜRo, Technik Und Wissenschaft (Btw), 9. Gi-Fachtagung (2001)
23. Stonebraker, M., Frew, J., Gardels, K., Meredith, J.: The Sequoia 2000 Benchmark. In: SIGMOD Conference, pp. 2–11 (1993)
24. O'Neil, P.E.: Database Performance Measurement. In: The Computer Science and Engineering Handbook, pp. 1078–1092. CRC Press (1997)
25. OLAP Council: APB-1 OLAP Benchmark Release II (1998), `http://www.olapcouncil.org/research/bmarkly.html`
26. Transaction Processing Performance Council (2000), `http://www.tpc.org`
27. Chaudhri, A.B.: Benchmarks (2000), `http://www.soi.city.ac.uk/~akmal/html.dir/benchmarks.html`
28. Tucker, P.A., Tufte, K., Papadimos, V., Maier, D.: Nexmark - a Benchmark for Querying Data Streams. Technical Report, OGI School of Science & Engineering at OHSU (2003)

29. Schmidt, A., Waas, F., Kersten, M., Carey, M.J., Manolescu, I., Busse, R.: XMark: a Benchmark for XML Data Management. In: Proc. of the 28th International Conference on Very Large Data Bases (2002)
30. Koch, C., Re, C., Olteanu, D., Lenz, H.J., Keulen, M.V., Haas, P.J., Pan, J.Z.: Working Group: Report of the Probabilistic Databases Benchmarking. In: Proc. of Dagstuhl Seminar 08421 on Uncertainty Management in Information Systems, pp. 12–17 (2008)
31. Arasu, A., Babu, S., Widom, J.: The CQL Continuous Query Language: Semantic Foundations and Query Execution. The VLDB Journal 15(2), 121–142 (2006)
32. Mitzenmacher, M., Upfal, E.: Probability & Computing: Randomized Algorithms and Probabilistic Analysis. Cambridge U. Press (2005)

Mining Same-Taste Users with Common Preference Patterns for Ubiquitous Exhibition Navigation

Shin-Yi Wu[1] and Li-Chen Cheng[2]

[1] Industrial Technology Research Institute, Hsinchu, Taiwan 310, ROC
[2] Department of Computer Science and Information Management, Soochow University, Taipei, Taiwan 100, ROC
sywu@itri.org.tw, lijencheng@csim.scu.edu.tw

Abstract. In a ubiquitous exhibition, an intelligent navigation service that can provide booths' information, recommend interesting booths and plan touring path is required for both visitors and vendors. The preference mining module is the kernel. This paper proposes a group-based user preference pattern mining method, which can be implemented as a preference mining module in this service. When the visiting traces that imply the preference of users are recorded, the method discovers user preference patterns with high representativeness and high discrimination from the historical visiting logs. According to the discovered model, collaborative recommendation can be accomplished, and then the intelligent navigation service can plan personalized touring path based on the recommendation lists. For demonstrating the performance of the proposed method, we engage some experiments, and then indicate the characteristics of the proposed method.

Keywords: Ubiquitous exhibition, user preference pattern, clustering, collaborative recommendation, data mining.

1 Introduction

Ubiquitous network society aims to create a convenient environment where people can reach any service via any ICT devices anytime anywhere [1]. Ubiquitous exhibition is a typical example. The meetings, incentives, conventions, and exhibitions (MICE) industry has been proven its economic impacts. Due to the globalized competition, turning international exhibitions into ubiquity is necessary [2]. In a ubiquitous environment, an intelligent navigation service that can lead visitors into as more booths they are really interested in as possible benefits both vendors and visitors. The main challenge of the intelligent navigation service is to predict the level of interests of each booth for a visitor.

The environment of an ideal ubiquitous exhibition could be illustrated as Fig. 1. The blocks marked A to W are booths. Each visitor may carry a personal mobile device along with him/her. The visiting behaviors of these visitors are logged and can be analyzed. Through the personal mobile device, a visitor may get personalized exhibition navigation service and the information about every booth anytime anywhere in this exhibition.

J.-S. Pan, S.-M. Chen, N.T. Nguyen (Eds.): ACIIDS 2012, Part III, LNAI 7198, pp. 416–425, 2012.

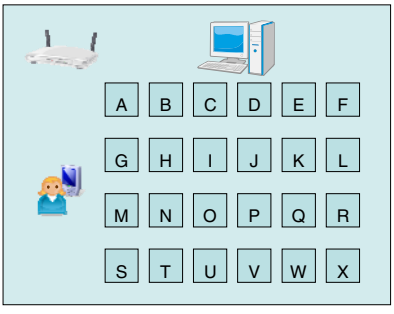


Fig. 1. An illustration of the ubiquitous exhibition environment

In a personalized exhibition navigation service, the following modules should be included. a) a wireless communication module that enables any visitor to access any service in a ubiquitous exhibition. b) a mobile trace module that identifies the locations (by any indoor positioning technology) and booth staying durations of a visitor and logs these visiting behaviors while he/she is visiting this exhibition. c) a preference mining module that analyzes the visiting logs of these visitors to discover some relevance among these visitors. d) a recommendation module that predicts the level of interests of booths for a visitor based on the result of the preference mining module, and outputs a recommendation list. e) an exhibition navigation module that navigates a visitor according to the recommendation list and the visiting trace of this visitor so far. The purpose of this paper is to propose a group-based user preference pattern mining method, which can be implemented as a preference mining module in a personalized exhibition navigation service.

In the next section, some works related to the domain of personalized recommendation in a ubiquitous exhibition are discussed. Section 3 formally defined the mining problem. In Section 4, we propose a method for this problem domain. Some experimental results are shown in Section 5. The conclusion and future works are given in Section 6.

2 Related Works

The traditional personalized recommendation methods could not fulfill the requirements in the ubiquitous exhibition scenario. Content-based methods [3], [4], [5], [6] are mostly designed for recommending text-based items. One major limitation of content-based methods is the capability to recommend dissimilar items. Collaborative filtering [7], [8], [9], [10], [11] is a popular method to predict user rating for item recommendation. Collaborative filtering predicts the active user's rating for an item based on this item ratings of his/her similar neighbors (the previous ratings between these users and the active user are similar). The sparsity problem is one major problem. Further, it can not provide the information how users are aggregated by their similar behaviors, and can not tell us the behavior patterns shared

by each group of similar users. These requirements are quite important for the recommendation service providers such as exhibition hosts and mobile device providers, because they want to earn advertisement profits from booth venders by providing a good personalized recommendation service. The frequent pattern-based clustering methods in the domain of clustering high-dimensional data are promising, such as frequent term-based text clustering [12], and *pCluster* [13]. Unfortunately, these methods will encounter problem when handling the even huge, sparse, or local centralized datasets which is common in user preference pattern mining scenario.

3 Problem Definition

Given a sequence database with three attributes: uid (user identity), eid (event identity), and duration (event staying duration), the mining method should distinguish the users into groups by these logged user sequences and find the commonly occurring patterns among users in the same group. For example, in the ubiquitous exhibition application, the sequence database can be transformed from the historical visiting logs of exhibition visitors. The uids are used to distinguish different visitors, an eid stands for a booth, and the duration is how long a visitor has stayed near a booth. Table 1 is a sample database.

Table 1. User sequence database *D*

uid	eid	du.	uid	eid	du.	uid	eid	du.	uid	eid	du.
1	0	1	3	1	1	6	0	8	8	8	1
1	1	10	3	2	1	6	2	4	8	9	2
1	5	1	4	0	2	6	3	5	9	3	13
1	2	1	4	1	16	6	8	9	9	4	1
1	6	13	4	5	1	6	9	13	9	7	5
1	0	6	4	2	1	6	0	4	9	8	2
1	1	2	4	0	5	6	1	7	9	4	1
2	0	1	4	1	1	6	6	1	10	3	16
2	1	8	4	6	10	7	0	10	10	4	2
2	5	1	5	0	10	7	2	4	10	7	5
2	6	10	5	2	7	7	3	5	10	8	1
2	1	2	5	3	5	7	8	10	10	0	6
3	0	1	5	8	8	7	0	4	10	3	2
3	1	8	5	9	4	8	3	11	10	6	2
3	5	2	5	0	4	8	4	1			
3	6	11	5	6	1	8	7	5			

In this mining problem, we assume that the length of duration is positively related to the degree of interest the user shows in that event. Taking the ubiquitous exhibition application as an example, if a customer stays around booth A longer than around booth B, it always means that he/she has more interests in booth A than booth B. Based on this assumption, the staying duration can be transformed to interesting degrees. For example, we can set the number of duration types (δ) as 3, which means

"like", "average", and "dislike"; as 5, which means "very like", "like", "average", "dislike" and "very dislike". If we set $\delta = 3$, the corresponding duration upper bounds and duration lower bounds should be sets of three values, respectively. In the database in Table 1, we may set the duration lower bounds as (1, 4, 8) and the duration upper bounds as (3, 7, ∞). Thus, the first three events in data sequence with uid = 1 will be transformed into (0, 0), (1, 2), (5, 0), respectively.

In the group-based user preference pattern mining problem, the purpose is to group users according to the *similarity* among them, and the common patterns among the users in each group should be indentified. The definition of similarity is defined in Definitions 1 and 2, and Example 1 explains these definitions. Besides, to ensure the discovered patterns are discriminative and representative, it is necessary to compute two *supports* of patterns, *support within groups* (Sup^{in}) and *support between groups* (Sup^{be}), which have been defined in Definitions 3, and 4.

Definition 1. (Similarity) Let E be the set of all possible events in user sequence database D and $DT = \{d_1, d_2, .., d_\delta\}$ be the set of δ duration types. Given two objects (users), $u_a = \{(e_1^a, dt_1^a), (e_2^a, dt_2^a), \ldots, (e_m^a, dt_m^a)\}$, $u_b = \{(e_1^b, dt_1^b), (e_2^b, dt_2^b), \ldots, (e_n^b, dt_n^b)\}$, where $e_1^a, .., e_m^a$ and $e_1^b, .., e_n^b \in E$, and $dt_1^a, .., dt_m^a$ and $dt_1^b, .., dt_n^b \in DT$. The similarity of u_a and u_b is defined as equation (1).

$$Sim(u_a, u_b) = w_{d_1} * Sim_{d_1}(u_a, u_b) + w_{d_2} * Sim_{d_2}(u_a, u_b) + \ldots + w_{d_\delta} * Sim_{d_\delta}(u_a, u_b), \quad \text{where } w_{d_1} + w_{d2} + \ldots + w_{d_\delta} = 1 \tag{1}$$

In equation (1), w_{d_1}, w_{d2}, …, w_{d_δ} are weights for each duration type. Since the significance of each duration type may be different, the weights can be set by applications. $Sim_{d_i}(u_a, u_b)$, $i \in 1, .., \delta$, is the similarity of u_a and u_b about the dimension of duration type d_i, which is defined in Definition 2.

Definition 2. (Similarity for duration type d_i) Following to Definition 1, let B_a^i and B_b^i be the set of occurring events with duration type d_i in sequences u_a and u_b, respectively. $|B_a^i|$ and $|B_b^i|$ are the numbers of items in sets B_a^i and B_b^i, respectively. $Sim_{d_i}(u_a, u_b)$ is defined as equation (2).

$$Sim_{d_i}(u_a, u_b) = \frac{2 * |B_a^i \cap B_b^i|}{|B_a^i| + |B_b^i|}, \quad i = 1 \cdots \delta \tag{2}$$
$$\text{If } |B_a^i| \models 0 \; and \; |B_b^i| \models 0, \quad Sim_{d_1}(u_a, u_b) = 0.5$$

Example 1 (Similarity) In Table 1, according to the above settings about duration, , sequences with uid = 1 (u_1) and with uid = 2 (u_2) are transferred as u_1 = {(0, 0), (5, 0), (2, 0), (1, 0), (0, 1), (1, 2), (6, 2)} and u_2 = {(0, 0), (5, 0), (1, 0), (1, 2), (6, 2)}. In this example, duration type set DT = {0, 1, 2}, which means {short, medium, long}.

According to equation (1), $Sim(u_1, u_2) = 0.4 * Sim_0(u_1, u_2) + 0.2 * Sim_1(u_1, u_2) + 0.4 * Sim_2(u_1, u_2)$, if we set $w_1 = 0.4$, $w_2 = 0.2$, and $w_3 = 0.4$. In order to compute $Sim_0(u_1, u_2)$, $Sim_1(u_1, u_2)$, and $Sim_2(u_1, u_2)$, the events in u_1 and u_2 are separated by their duration types. First, we compute $Sim_0(u_1, u_2)$. Since there are four events with short duration type in u_1(B_1^0), and three in u_2 (B_2^0), and the intersection of B_1^0 and B_2^0 are events 0, 5, and 1 (three events), $Sim_0(u_1, u_2) = 2 * 3 / (4 + 3) = 0.857$ according to equation (2). By the same way, we may compute $Sim_1(u_1, u_2) = 0$ and $Sim_2(u_1, u_2) = 1$. Thus, $Sim(u_1, u_2) = 0.4 * 0.857 + 0.2 * 0 + 0.4 * 1 = 0.743$.

Definition 3. (Support within group) Support within group (denoted as Sup^{in}) is used to measure how representative an event is in its cluster (i.e. group). A simple way to measure the Sup^{in} of event e in cluster c is defined in equation (3).

$$Sup^{in}(e, c) = \frac{|\, u \mid e \text{ is existed in the event sequences associated with user } u \,|}{|\, u \mid u \in c \,|} \tag{3}$$

Example 2 (Support within group) In Table 1, assume that there are five users, in cluster c which are with uid = 1, 4, 8, 9, and 10, respectively. Let events $e_1 = (0, 0)$, $e_2 = (0, 1)$, and $e_3 = (9, 0)$. According to equation (3), we may compute $Sup^{in}(e_1, c) = 2 / 5 = 0.4$, $Sup^{in}(e_2, c) = 3 / 5 = 0.6$, and $Sup^{in}(e_3, c) = 1 / 5 = 0.2$.

Definition 4. (Support between groups) Support between groups (denoted as Sup^{be}) is used to measure the discrimination of an event between clusters. The higher Sup^{be} it is, the lower discrimination the event is. Let C be the set of all user clusters for database D and there are k clusters in C, a simple way to measure the Sup^{be} of event e in cluster c is defined in equation (4).

$$Sup^{be}(e) = \frac{|\, c \mid e \text{ is one of event in the pattern of cluster } c, \forall c \in C \,|}{k}, \tag{4}$$

Example 3 (Support between groups) Assume that the data in Table 1 are grouped into three clusters c_1, c_2, and c_3, and the patterns are $p_1 = \{(0, 1), (4, 0), (8, 0), (7, 1), (3, 2)\}$, $p_2 = \{(0, 0), (5, 0), (1, 0), (1, 2), (6, 2)\}$, and $p_3 = \{(6, 0), (2, 1), (3, 1), (9, 1), (0, 1), (0, 2), (8, 2)\}$, respectively. According to equation (4), we may compute $Sup^{be}((0, 1)) = 0.667$, $Sup^{be}((4, 0)) = 0.333$, and $Sup^{be}((3, 2)) = 0.333$.

The purpose of the group-based user preference pattern mining problem is to satisfy the following three criteria:

1. Making the highest similarity among objects in the same cluster as possible.
2. Given a threshold min_sup^{in}, the Sup^{in} of every user preference pattern event should be greater than or equal to min_sup^{in}.
3. Given a threshold max_sup^{be}, the Sup^{be} of every user preference pattern event should be less than or equal to max_sup^{be}.

4 Mining User Preference Patterns

To solve the mining problem defined above, we develop an algorithm *CUP* (Clustering and discovering User Preference pattern). *CUP* is designed by modified the famous clustering method *k*-means [14] for handling data that are high-dimensional and include many missing values. In addition, the discrimination and representativeness were taken into considerations.

Method: Call CUP(k, D, min_sup^{in}, max_sup^{be}, min_sim, δ, Du_UBound_list, Du_LBound_list, weight_list)
Input: *k*: The maximum number of user clusters; *D*: User sequence database; min_sup^{in}: Minimum support within group; max_sup^{be}: Maximum support between groups; min_sim: Minimum similarity (optional); δ: The number of duration types (optional); Du_UBound_list: A list of δ values for duration upper bounds (optional); Du_LBound_list: A list of δ values for duration lower bounds (optional); weight_list: A list of δ values for weights of all duration type (optional).
Output: *c*: clusters with user preference patterns.

```
01    Procedure CUP(k, D, min_sup^in, max_sup^be, min_sim, δ,
      Du_UBound_list, Du_LBound_list, weight_list) {
02    Q = φ; //Candidate center queue (maintained
      ascending by object's similarities)
03    Arbitrary choose k objects from D as the initial
      cluster centers;
04    Repeat
05
06        If Q ≠ φ
07          For each cluster c_i
              If the pattern length of c_i is 0 or the
              number of objects in a cluster <= 1, then
08            pop an object from Q as the center of this
09            cluster;
            Q = φ;
          (Re)assign each object o_i to the cluster c_j to
10        which o_i is the most similar, based on the
11        cluster patterns (centers);
            If the largest similarity for o_i is lower
12          than min_sim, then push o_i into Q;
13        Generate user preference pattern under min_sup^in
          threshold for each cluster as its new center;
14    Until no change;
15    Filtering indiscriminative events from preference
      patterns by max_sup^be;
      Output clusters with preference patterns.
      }
```

The first step of *CUP* is randomly choosing an object as the center for each cluster (line 3). The center of a cluster is an actual object in the first run of clustering, but in other runs, the center of a cluster is its pattern. From line 4 to line 12, it will do several runs to find a best clustering result. After the centers of clusters are selected, every object will be assigned into the closed cluster (line 9). If the similarity between an object and a cluster center is highest among all clusters, the object will be assigned to this closest cluster. Once all objects are assigned into clusters, *CUP* starts to compute Sup^{in} for each event in each cluster (line 11). In cluster c, if the Sup^{in} of an event is greater than or equal to min_sup^{in}, e will be added into the pattern of c (line 11). The patterns, however, are candidates, because they may not satisfy the min_sup^{in} threshold. These patterns become the cluster centers in the next iteration. The loops will keep running until the objects don't move. After the loop, *CUP* then computes Sup^{be} for each pattern and filters events in patterns by max_sup^{be} (line 14). In Example 4, we give a running example of *CUP*.

In the above process, we found that once a cluster contained no objects, it will be empty forever. It is due to a bad initial center selection or inadequate threshold settings (too harsh). With a bad cluster center selection the objects assigned in this cluster may be dissimilar. Thus the user preference pattern may not be generated correctly (maybe empty). After that, this cluster will never been assigned any objects. Center replacing could resolve this problem. An object dissimilar to all clusters (an outlier) should be a good candidate to replace a bad cluster center. We, consequently, implement a queue to collect these candidates, and then replace the empty clusters by the objects in this queue. Lines 2, 6, 7, 8, and 10 state the above ideas.

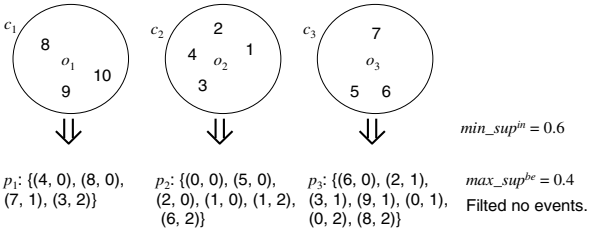


Fig. 2. Final results for the running example

Example 4 (Running example for CUP) The input parameters are set as below: k = 3, D = the user sequence database in Table 1, min_sup^{in} = 0.6, max_sup^{be} = 0.4, δ = 3, *Du_UBound_list* = (3, 7, ∞), *Du_LBound_list* = (1, 4, 8), and *weight_list* = (0.4, 0.2, 0.4). According to this setting, data sequences in D could be transformed. For example, user sequence of user 1 and user 2 were transformed into {(0, 0), (5, 0), (2, 0), (1, 0), (0, 1), (1, 2), (6, 2)} and {(0, 0), (5, 0), (1, 0), (1, 2), (6, 2)}, respectively. Assume the initial centers are users 1, 2, and 7. After computing the similarities, the objects will be assigned as c_1 = {user1, user4, user 8, user 9, user 10}, c_2 = {user2, user3}, and c_3 = {user5, user6, user 7}. Since min_sup^{in} = 0.6, the cluster patterns can be generated. The pattern p_1 of cluster c_1 is {(4, 0), (8, 0), (0, 1), (7, 1), (3, 2)}, the

pattern p_2 of cluster c_2 is {(0, 0), (5, 0), (1, 0), (1, 2), (6, 2)}, and the pattern p_3 of cluster c_3 is {(6, 0), (2, 1), (3, 1), (0, 1), (9, 1), (0, 2), (8, 2)}. These three patterns become the cluster centers in the next run. In this example, it takes three runs to get the final result (Fig. 2). Since there is not any events which violates $max\_sup^{in} = 0.4$, no events are filtered here.

5 Experiments

For verifying the efficiency of the proposed method, we develop a data generator to generate synthetic data to test the run time under different data configurations and to test the scalability. The data generator is designed by simulating the user visiting logs in a ubiquitous exhibition. The parameters are: a) $|D|$: The number of users; b) L_u: Average number of items for each user; c) L_p: Average number of items for potentially large patterns; d) N: The number of events; e) k_E: The user expected number of clusters.

To test the efficiency of *CUP*, we use the data generator to generate six datasets with configuration settings listed in Table 2. k_E is fixed to 10. When executing *CUP* on these six datasets, the threshold min_sup^{in} is set to 0.4, 0.6, and 0.8, and max_sup^{be} is set to 0.3 and 0.5. The other parameters for *CUP* are fixed as $k = 3$, $min\_sim = 0.1$, $\delta = 3$, $Du\_LBound\_list = (1, 4, 8)$, $Du\_UBound\_list = (3, 7, \infty)$, and $weight\_list = (0.4, 0.2, 0.4)$. In real cases, the parameters δ, *Du_LBound_list*, and *Du_UBound_list* could better be decided according to the data distribution.

Table 2. Configuration settings

	$\|D\|$	L_u	L_p	N
D1-Lu50-Lp15-N0.1	1,000	50	15	100
D2-Lu50-Lp15-N0.1	2,000	50	15	100
D1-Lu100-Lp15-N0.1	1,000	100	15	100
D1-Lu50-Lp30-N0.1	1,000	50	30	100
D1-Lu100-Lp30-N0.1	1,000	100	30	100
D1-Lu50-Lp15-N0.2	1,000	50	15	200

The experiment results are shown in Fig. 3. The six lines show the execution time of *CUP* on dataset of different configuration settings in Table 2. When reading Fig. 3, we can take the dark blue line (D1-Lu50-Lp15-N0.1) as the base setting. For example, because the pink line doubles the number of users ($|D|$), *CUP* spent more time in mining this dataset than the base one. One may notice that the execution time taken on D1-Lu50-Lp15-N0.1 (blue line) at the support setting (0.8, 0.3) is exceptionally longer than others. This is reasonable because in this support setting, it takes more iterations to mine results. In some configurations such as D1-Lu50-Lp30-N0.1 (light blue) and D1-Lu50-Lp15-N0.2 (red), *CUP* spent less time to execute than base setting, thus we know that the effects of L_p and N may not significant, since it may take less iterations to accomplish mining. The information about number of iterations can refer to Fig. 4. In this experiment, some support settings may be too strict, so

CUP can not get results in 50 iterations. This is why some points are missed in these diagrams. In general, the execution time evaluation for *CUP* is quite satisfactory, because each of six datasets spent *CUP* no more than 3.5 seconds.

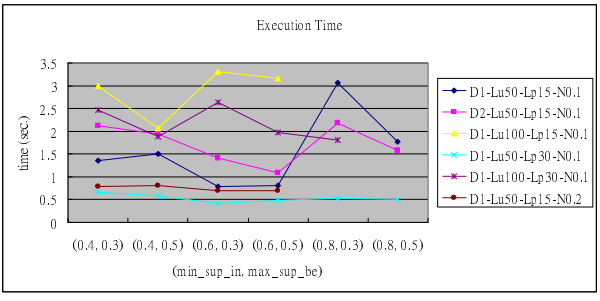


Fig. 3. Efficiency testing results

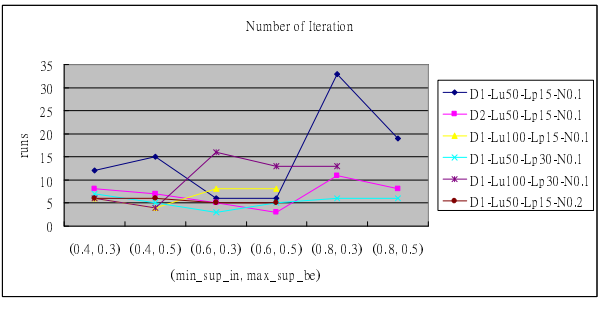


Fig. 4. Number of iteration for efficiency testing results

In real applications, the scalability of a mining method is important. We, then, do scalability evaluations for *CUP* on testing the scalability for |*D*| and *N*. In both evaluations, we use the base setting D1-Lu50-Lp15-N0.1 but vary |*D*| and *N*. Support settings are fixed to (min_sup^{in}, max_sup^{be}) = (0.6, 0.5). The experiment results shown that the effect to execution time by |*D*| is approximately polynomial, and the effect to execution time by *N* is almost linear.

6 Conclusion

In this paper, we claim that the previous recommendation methods could not satisfy the requirements for making personalized recommendation in the ubiquitous exhibition and many other applications. A new mining problem, group-based user preference pattern mining, is defined, and then a novel algorithm *CUP* is proposed to solve this problem. *CUP* is verified to have good performance on mining synthetic datasets. The mining results have shown that group-based user preference pattern mining technique can satisfy the two requirements the above claimed.

This work can be extended in several directions. First, one valuable improvement is to solve the local optimal problem of *CUP*, which is inherited from *k*-means. Second, to adopt *CUP* in real datasets to evaluate its effectiveness is important, too. Third, in order to make recommendation in real applications, it is necessary to propose a framework considering all possible situations such as the new item problem and the new user problem.

References

1. Lee, S.H., Han, J.H., Leem, Y.T., Yigitcanlar, T.: Towards Ubiquitous City: Concept, Planning, and Experiences, pp. 148–169. Igi Global (2008)
2. Lua, T., Fub, L.: Toward Full Coverage UHF RFID Services-An Implementation in Ubiquitous Exhibition Service. In: Proceedings of the 16th ISPE International Conference on Concurrent Engineering, pp. 501–510. Springer, Heidelberg (2009)
3. Lang, K.: Newsweeder: Learning to filter netnews. In: Proc. 12th Int'l. Conf. Machine Learning (1995)
4. Mooney, R.J., Roy, L.: Content-based book recommending using learning for text categorization. In: Proc. ACM SIGIR 1999 Workshop Recommender System: Algorithms and Evaluation, pp. 195–204. ACM, New York (2000)
5. Pazzani, M., Billsus, D.: Learning and revising user profiles: The identification of interesting web sites. Machine Learning 27, 313–331 (1997)
6. Balabanovic, M., Shoham, Y.: Fab: Content-based, collaborative recommendation. Comm. ACM 40, 66–72 (1997)
7. Herlocker, J.L., Konstan, J.A., Riedl, J.: An algorithmic framework for performing collaborative filtering. In: Proc. 22nd Int'l ACM SIGIR Conf. Research and Development in Information Retrieval, pp. 230–237. ACM, New York (1999)
8. Konstan, J.A., Miller, B.N., Maltz, D., Herlocker, J.L., Gordon, L.R., Riedl, J.: GroupLens: applying collaborative filtering to Usenet news. Comm. ACM 40, 77–87 (1997)
9. Terveen, L., Hill, W., Amento, B., McDonald, D., Creter, J.: PHOAKS: A system for sharing recommendations. Comm. ACM 40, 59–62 (1997)
10. Hill, W., Stead, L., Rosenstein, M., Furnas, G.: Recommending and evaluating choices in a virtual community of use. In: Proc. Conf. Human Factors in Computing Systems, pp. 194–201. ACM Press/Addison-Wesley Publishing Co., New York, USA (1995)
11. Goldberg, D., Nichols, D., Oki, B.M., Terry, D.: Using collaborative filtering to weave an information tapestry. Comm. ACM 35, 61–70 (1992)
12. Beil, F., Ester, M., Xu, X.: Frequent term-based text clustering. In: Proc. 8th Int. Conf. on Knowledge Discovery and Data Mining (KDD), Alberta, Canada, pp. 436–442 (2002)
13. Wang, H., Wang, W., Yang, J., Yu, P.S.: Clustering by pattern similarity in large data sets. In: Proc. 2002 ACM-SIGMOD Int. Conf. on Management of Data (SIGMOD 2002), pp. 394–405 (2002)
14. MacQueen, J.: Some methods for classification and analysis of multivariate observations. In: Proc. Fifth Berkeley Symp. on Mathematical Statistics and Probability, vol. 1, pp. 281–297. Univ. of Calif. Press (1966)

Publication Venue Recommendation Using Author Network's Publication History

Hiep Luong[1], Tin Huynh[2], Susan Gauch[1], Loc Do[2], and Kiem Hoang[2]

[1] University of Arkansas, U.S.A.
{hluong,sgauch}@uark.edu
[2] University of Information Technology, Vietnam
{tinhn,locdo,kiemhv}@uit.edu.vn

Abstract. Selecting a good conference or journal in which to publish a new article is very important to many researchers and scholars. The choice of publication venue is usually based on the author's existing knowledge of publication venues in their research domain or the match of the conference topics with their paper content. They may not be aware of new or other more appropriate conference venues to which their paper could be submitted. A traditional way to recommend a conference to a researcher is by analyzing her paper and comparing it to the topics of different conferences using content-based analysis. However, this approach can make errors due to mismatches caused by ambiguity in text comparisons. In this paper, we present a new approach allowing researchers to automatically find appropriate publication venues for their research paper by exploring author's network of related co-authors and other researchers in the same domain. This work is a part of our social network based recommendation research for research publications venues and interesting hot-topic researches. Experiments with a set of ACM SIG conferences show that our new approach outperforms the content-based approach and provides accurate recommendation. This works also demonstrates the feasibility of our ongoing approach aimed at using social network analysis of researchers and experts in the relevant research domains for a variety of recommendation tasks.

Keywords: recommender systems, publication history, kNN, machine learning, social network analysis.

1 Introduction

With the enormous growth and complexity of information that is added to the Web daily, it is a challenging task for users to find exactly what they are looking for or for researchers to keep up to date on information of whose existence they may be unaware. Recommender systems are one solution to helping users deal with the flood of information. They are tools that automatically filter a large set of items, e.g., movies, books, scientific papers, music, etc., in order to identify those that are most relevant to a user's interest. Recommender systems basically are divided into three categories: (1) content-based filtering; (2) collaborative

J.-S. Pan, S.-M. Chen, N.T. Nguyen (Eds.): ACIIDS 2012, Part III, LNAI 7198, pp. 426–435, 2012.

filtering and (3) hybrid recommendation systems. The content-based filtering uses actual content features of items, while the collaborative filtering predict new user's preference using other users' rating, assuming the like-minded people tend to have similar choices [4].

Recommendation systems are particularly important for researchers and scholars in their professional research activities. For some experts in a research domain, or senior researchers who have strong publication records, selecting a conference might be a trivial task since they know well which conferences, journals or scientific forums are the best places in which to publish their research papers. However, many other people who have less publication experience or are not current on new publication venues in their research domain. For these researchers, an automatic system that recommends relevant venues matching their profiles, their professional networks and research interests could be particularly useful.

In this work, we present a new method to recommend conferences, journals for researchers to submit their paper based on social networking and analyzing for researchers in the computer science area. Our approach is empirically evaluated using a dataset of recent ACM conference publications and compared with a baseline content-based.

2 Related Work

Automated text categorization is considered as the core of content-based recommendation systems. This supervised learning task is defined as assigning category labels (pre-defined) to new documents based on their likelihood of belonging to a given class as represented by a training set of labeled documents [16]. Yang and Liu reported a controlled study with statistical significance tests on five text categorization methods: Support Vector Machines (SVM), k-Nearest Neighbor (kNN) classifier, Neural Network approach, Linear Least-squares Fit mapping and a Naive Bayes classifier [16]. Their experiments with the Reuters data set showed that SVM and kNN significantly outperform the other classifiers, while Naive Bayes underperforms all the other classifiers. In other work [6], kNN was found to be an effective and easy to implement that could, with appropriate feature selection and weighting, outperforms SVM [6].

The online world supported the creation of many research-focused digital libraries such as the Web of Science, ACM Portal, Springer Link, IEEE Xplore, Google Scholar, and CiteSeer^X. Initially, these were viewed as somewhat static collections of research literature. However, recent research is focusing on these as representations of a community of scholars, building and analyzing social networks of researchers to extract useful information about research domains, user behaviors, and the relationships between individual researchers and the community as a whole. Microsoft Academic Search [1], ArNetMiner [15], and AcaSoNet [3] are online, web-based systems whose goal is to identify and support communities of scholars via their publications. The entire field of social network systems for the academic community is growing quickly, as evidenced by the number of other approaches being investigated [2] [13] [5] [11] [12].

Social network analysis (SNA) is a quantitative analysis of relationships between individuals or organizations to identify most important actors, group formations or equivalent roles of actors within a social network [9]. SNA is considered a practical method to improve knowledge sharing [14] and it is being applied in a wide variety of contexts. [10] and [8] apply SNA to enhance an information retrieval (IR) systems. To our best knowledge, we have not seen existing research that applies social network analysis to recommend appropriate publication venues.

3 Our Approach

In this section, we present two main approaches to recommending a list of appropriate conference venues to an author for their unpublished paper. The first approach, a content-based recommendation, is considered as the baseline approach. We then introduce our new approach by using the social network analysis to explore author network's publication history in order to generate relevant conferences venues.

3.1 Overview

There are a wide variety of dissemination outlets for research results, e.g., conferences, journals, seminars, scientific forums. When an author has a paper that they want to share, the review cycle can be time consuming and, if the paper is rejected because it is not a good fit, valuable time can be lost. In computer science, in particular, the pace of innovation is high. Selecting the right publication venue the first time is particularly important.

To prepare data for our experiment, we selected four subdomains of research in Computer Science corresponding to four SIGs (Special Interest Groups). Then, we manually picked four different conferences for each SIG. The total list of selected conferences contained 16 conferences, which is listed as following:

1. SIGBED - Special Interest Group on Embedded System
   CASES - Compilers, Architecture, and Synthesis for Embedded Systems
   CODES+ISSS - International Conference on Hardware/Software Codesign and Systems Synthesis
   EMSOFT - International Workshop on Embedded Systems
   SENSYS - Conference On Embedded Networked Sensor Systems
2. SIGDA - Special Interest Group on Design Automation
   DAC - Design Automation Conference
   DATE - Design, Automation, and Test in Europe
   ICCAD - International Conference on Computer Aided Design
   SBCCI - Annual Symposium On Integrated Circuits And System Design
3. SIGIR - Special Interest Group on Information Retrieval
   CIKM - International Conference on Information and Knowledge Management
   JCDL - ACM/IEEE Joint Conference on Digital Libraries
   SIGIR - Research and Development in Information Retrieval
   WWW - World Wide Web Conference Series
4. SIGPLAN - Special Interest Group on Programming Languages
   GPCE - Generative Programming and Component Engineering
   ICFP - International Conference on Functional Programming
   OOPSLA - Conference on Object-Oriented Programming Systems, Languages, and Applications
   PLDI - SIGPLAN Conference on Programming Language Design and Implementation

We have downloaded all published papers of these 16 conferences in three recent years, i.e. 2008-2010, from the ACM digital library[1] . Since we know already the correct conferences for these papers, we can this information as the truth-list to evaluate our conference recommendation approaches.

[1] http://dl.acm.org/

3.2 Baseline: Content-Based Recommendation Approach

Content-based classification has been widely used in recommendation applications. One of the most commonly used algorithms in recommender systems is the k-nearest neighborhood (kNN) approach. We have applied the kNN algorithm [7] in implementing the conference recommendation system based on papers' content in the collected dataset.

We first built a conference classification that attempts to predict the correct conference for a paper based on its content. First, we trained the kNN classifier using some documents selected from the papers published in 2008 and 2009. Then, the papers published in 2010 were used to evaluate the classifier.

3.3 Author Network Analysis Approach

We developed a new approach that builds a social network for each author and then recommends conferences based on the reputation of the author's social network and other information such as conference name, conference's subdomain, number of publications.

For each paper in the dataset, we extracted the author names and used a crawler to gather information about each author's publications and co-authors from Microsoft Academic Search[2] website (MAS). The co-authors of the co-authors were then recursively collected, until a network of 3 levels deep was created. In this paper, we present three methods to exploit this social network information in a recommendation system.

Method 1: Most frequent conference

In the first method, we create a social network for each candidate paper by combining the social networks for each co-author of the paper. We then recommend the conference that these authors have published in most often in the past.

We weight the candidate conference values by simply calculating the total number of papers published to that conference by authors in the network.

$$freq\_CONF_i = \sum_{j=1}^{N} num\_paper_{i,j} \tag{1}$$

where N is the total number of authors having published paper in the conference i. Suppose we are building a conference recommendation system in which the input includes four conferences, i.e., SIGIR, CIKM, KDD and WWW. We need to find the most appropriate conference for a given test paper whose main authors are Alex and Paul. First, we use MAS to extract Alex and Paul's co-authors on other papers. In this example, we build a network of 2 authors (Alex and Paul) and their 3 co-authors (David, Adam, and Josh). Then, we extract the publication history for each person in the network, as presented in the Table 1.

We count the total number of papers for each conference listed in the network; the corresponding values are as presented in the Table 1. (i.e., column Method

[2] http://academic.research.microsoft.com/

Table 1. Conferences and number of publications in author's network

Confs.	*Num. papers for each author in the network*					**Method1 (Total)**	*Frequent conference normalized by author*					**Method2 (Total)**
	David	**Adam**	**Josh**	**Alex**	**Paul**		**David**	**Adam**	**Josh**	**Alex**	**Paul**	
SIGIR	25	5	5	5	0	**40**	25/50	5/25	5/20	5/15	0	1.28
CIKM	10	10	5	5	9	39	10/50	10/25	5/20	5/15	9/10	**2.08**
KDD	10	0	5	0	0	15	10/50	0	5/20	0	0	0.45
WWW	5	10	5	5	1	26	5/50	10/25	5/20	5/15	1/10	1.18
Total	**50**	**25**	**20**	**15**	**10**		**1**	**1**	**1**	**1**	**1**	

1). In this example, SIGIR would be selected as the most likely conference for the paper.

Method 2: Most frequent conference normalized by author

The above method, since it merely counts the number of publications, could allow a highly published author to have more influence on the selection than the others. For comparison, we also implemented a method in which each author has equal influence on the outcome. Essentially, we calculate the prior probability for each author for each conference so that each author's total contribution sums to 1.0. We calculate the weight of a conference i as the sum of each author's probability of publication in conference i.

$$nfreq\_CONF_i = \sum_{j=1}^{N} \left(\frac{num\_paper_{i,j}}{total\_paper_j} \right) \tag{2}$$

Where N is the total number of authors having published paper in the conference i, $num\_paper_{i,j}$ is the number of papers that the author j has published at the conference i, and $total\_paper_j$ is her total number of paper published. Using the same example as above, we get weights calculated for each conference as presented in the column Method 2, Table 1. Based on this method, CIKM would be recommended as the most appropriate conference.

Method 3: Method 2 incorporating network topology

Both previous methods treat all authors the same, ignoring the strength of connections between the authors in the network. In this method, we not only take into account information about publications but also the co-author relationships among members of the network. We weight the contributions of each co-author by the number of papers they have co-authored with the main author. Thus, co-authors who publish with the main author frequently have more influence on the conference selected. Table 2 presents, for each author, their co-authors with the number of co-authored papers. For example, Alex has co-authored 15 papers with David, 2 with Adam, etc. The weight of a conference by this method is calculated as following:

$$coauthorship\_CONF_i = \sum_{m=1}^{A} coauthors\_w_{i,m} \quad (3)$$

where A is main author(s) of the test paper, $coauthors\_w_{i,m}$ is the co-authors' conference weight between the main author m and her co-authors in the network.

$$coauthors\_w_{i,m} = \sum_{k=1}^{CoA} (nfreq\_CONF_{i,m} + nfreq\_CONF_{i,k}) * w\_CoA_{k,m} \quad (4)$$

where CoA is co-author(s) of the main author m who have published respectively at the conference i, $w\_CoA_{k,m}$ is the number of times a main author m has co-authorships with other member k in the network. Since both Alex and Paul are main authors of the paper, we calculated the weight for each conference as the summarized value from each main author's conference scores. With the highest weight returned, i.e. 20.36, we recommend CIKM as the most relevant conference for the test paper.

Table 2. Co-authorships and conference weight calculation in the Method 3

Confs.	*Alex*				*Paul*				Method3 (Total)
	David (15)	Adam (2)	Josh (3)	Paul (1)	David (4)	Adam (0)	Josh (1)	Alex (1)	
SIGIR	(0.33+0.5)*15	(0.33+0.2)*2	(0.33+0.25)*3	(0.33+0)*1	(0+0.5)*4	(0+0.2)*0	(0+0.25)*1	(0+0.33)*1	18.16
CIKM	(0.33+0.2)*15	(0.33+0.4)*2	(0.33+0.25)*3	(0.33+0.9)*1	(0.9+0.5)*4	(0.9+0.2)*0	(0.9+0.25)*1	(0.9+0.33)*1	**20.36**
KDD	(0+0.2)*15	(0+0)*2	(0+0.25)*3	(0+0)*1	(0+0.2)*4	(0+0)*0	(0+0.25)*1	(0+0)*1	4.8
WWW	(0.33+0.1)*15	(0.33+0.4)*2	(0.33+0.25)*3	(0.33+0.1)*1	(0.1+0.1)*4	(0.1+0.4)*0	(0.1+0.25)*1	(0.1+0.33)*1	11.66

4 Experiments and Evaluation

4.1 Dataset

As presented in the section 3.1, our dataset contains 16 ACM conferences of 4 SIGs. With all papers collected from 2008-2010, we used the papers published in two years 2008 and 2009 as training documents and the ones published in 2010 as test documents for the classification task. Since the number of papers published for each conference varies, we randomly selected 20 documents per year (60 totals) for our dataset. Thus, each conference had 40 training and 20 test documents. With 16 conferences, the collection contains 640 training and 320 test documents. The test documents were further randomly divided into two subsets of 160, half for tuning our classifiers, and the other half for validation.

In order to get the publication history of authors, we developed a focused crawler in Java that extracts all co-authors and relevant publications for a given author from the Microsoft Academic Search[3] (MAS) website. This website updates currently more than 27 million of publications and 16 million of authors

[3] http://academic.research.microsoft.com/

with several thousand new updates every week in natural and life sciences. The crawler can be configured to recursively collect co-authors up to a selected depth. We ran the crawler for authors of all papers in the test and tuning data sets, collecting information about 115,000 authors and 160,000 papers. Because there is no tuning required in our network-based approach, we only ran the crawler to collect further information (co-authors, papers, conferences, etc.) for the authors in the test set. This information was downloaded and stored in a database.

4.2 Baseline Experiment for the Content-Based Approach

We applied the kNN classification module implemented in the KeyConcept package [7] to test our content-based approach. First, we trained the classifier with the 640 training documents. In order to identify the best k value for our experiment, we varied the k value, i.e. number of documents used for training, from 3-40. For each k value, we tested with the 160 tuning papers. Classification results, showing the accuracy of the top 3 predicted conferences, are reported in the Table 3.

Table 3. Classification results using the kNN algorithm

k	*#correct conferences found*			*Precision*		
	Top1	**Top2**	**Top3**	**Top1**	**Top2**	**Top3**
3	40	67	90	25.00%	41.88%	56.25%
5	57	91	121	35.63%	56.88%	75.63%
9	56	88	111	35.00%	55.00%	69.38%
12	63	95	116	39.38%	59.38%	72.50%
15	63	95	115	39.38%	59.38%	71.88%
18	70	97	117	43.75%	60.63%	73.13%
21	71	105	124	44.38%	65.63%	77.50%
25	**71**	**105**	**129**	**44.38%**	**65.63%**	**80.63%**
30	66	105	130	41.25%	65.63%	81.25%
40	71	111	128	44.38%	69.38%	80.00%

The best case for the content-based approach was achieved at k = 25, i.e., the number of documents returned by classifier used to predict the conference. We got the precision of 44.38%, 65.63%, and 80.63% for the Top1, Top2 and Top3 results. In order to confirm the stability of the content-based approach using kNN, we validated our results with k = 25 on the 160 validation papers. The validation test set actually produced slightly higher accuracy, with precision 48.75%, 68.75% and 80.00% for the Top1, Top2 and Top3 predicted conferences.

Since there are 16 conferences who overlap in topics, but only 4 SIGs that cover different research areas, predicting the correct SIG should be an easier task than predicting the correct conference. When we evaluate the kNN classifier on this task, we see higher accuracy of 80.63%, 92.50% and 94.38% for the Top1, Top2, and Top3 predicted SIG.

4.3 Author's Network Experiments

For each test paper, we built a network of main authors with their co-authors crawled from the MAS website. We then analyzed each main author's network to determine the author's publication frequencies in each conference. Among the 160 papers in the test set, there were 6 papers that were not found in the MAS website, so our results are reported over 154 total papers. Table 4 presents our classification results using author network's publication history using each of our three methods. We evaluated classification precision over two factors, accuracy in predicting the conference and accuracy predicting the SIG.

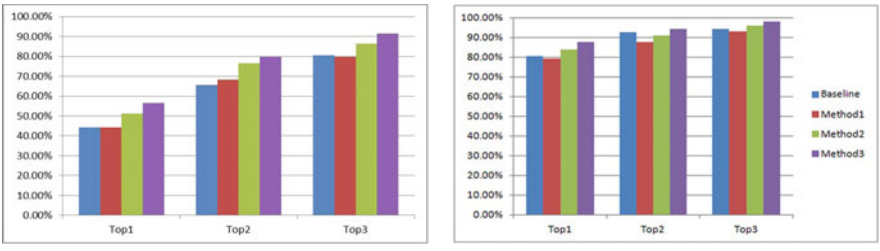
Table 4. Classification results using author networks

Method	*ACM conferences*						*ACM Special Interest Groups*					
	#correct found			*Precision*			*#correct found*			*Precision*		
	Top1	**Top2**	**Top3**	**Top1**	**Top2**	**Top3**	**Top1**	**Top2**	**Top3**	**Top1**	**Top2**	**Top3**
Method1	68	105	123	44.16%	68.18%	79.87%	122	135	143	79.22%	87.66%	92.86%
Method2	79	118	133	51.30%	76.62%	86.36%	129	140	148	83.77%	90.91%	96.10%
Method3	**87**	**123**	**141**	**56.49%**	**79.87%**	**91.56%**	**135**	**145**	**151**	**87.66%**	**94.16%**	**98.05%**
Baseline	***71***	***105***	***129***	***44.38%***	***65.63%***	***80.63%***	***129***	***148***	***151***	***80.63%***	***92.50%***	***94.38%***

4.4 Evaluation and Discussion

Figure 1 compares the accuracy of our three network-based publication history methods with the content-based approach. We found that the baseline method performed as well as Method 1, our simple frequency-based method. When we normalize the frequency method by author in Method 2, we outperform the baseline. Further, when we incorporate co-authorship relationships within the author network into the classifier in Method 3, we achieve the highest accuracy. Compared to the baseline (c.f., Figure 1(a)), Method 3 predicts the correct conference as the first choice 56.49% of the time (versus 44.38%) and within the first three results 91.56% of the time (versus 80.63%). In an easier test, identifying the correct SIGs, Method 3 predicted the correct SIG as its first result 87.66% of the time versus 80.63% for the baseline (c.f., Figure 1(b)).

Overall, even the simplest author network approach performs as well as the content-based approach, and the two more sophisticated author network approaches outperform the content-based approach. This is somewhat surprising because these approaches do not take into account anything about the paper contents themselves, just information about the social network of authors and their relationships to past conferences. On the other hand, content-based approaches have a difficult task deciding between multiple conferences that overlap on subject, requiring approaches with take into account more than just the keywords in the training and test documents. These results confirm our hypothesis that a publication venue recommendation system can benefit from social network analysis instead of, or in addition to, traditional content-based approaches.



(a) Conferences prediction accuracy (b) SIGs prediction accuracy

Fig. 1. Comparison of the publication venues prediction accuracy

5 Conclusions

The goal of this research is to implement and evaluate a new approach of using a social network of authors, linked by their publication history, to recommend publication venues for a new article. We presented three different methods to incorporate publication history into a classifier that predicts the appropriate conference. Our approach was empirically tested using a dataset of 16 ACM conferences from 4 SIGs and compared to a content-based approach using a kNN classification. The results show that our network-based methods outperform the content-based approach. Our best algorithm predicted the correct conferences within the top 3 conferences 91.56% of the time. In comparison, the content-based approach achieved 80.63% on the same task. If we compare the precision of the SIGs prediction, our methods also work better than the content-based method.

Our main tasks in the future are continuing to enhance the publication venues recommendation system by further exploiting the social network of authors, developing algorithms that take into account more sophisticated graph relationships, more nodes in the network (i.e., more levels of co-authorship relationships), and different kinds of links in the network (e.g., program committee membership). We will also develop a recommender system for new manuscripts and deploy this on the CiteSeer^X site.

Acknowledgments. This research is partially supported by the NSF grant number 0958123 - Collaborative Research: CI-ADDO-EN: Semantic CiteSeer^X.

References

1. Microsoft Academic Search, http://academic.research.microsoft.com
2. Abbasi, A., Altmann, J.: On the correlation between research performance and social network analysis measures applied to research collaboration networks. In: Proceedings of the 2011 44th Hawaii International Conference on System Sciences, HICSS 2011, pp. 1–10. IEEE Computer Society, Washington, DC (2011)

3. Abbasi, A., Altmann, J.: A social network system for analyzing publication activities of researchers. TEMEP Discussion Papers 201058, Seoul National University; Technology Management, Economics, and Policy Program, TEMEP (2010)
4. Adomavicius, G., Tuzhilin, A.: Toward the next generation of recommender systems: A survey of the state-of-the-art and possible extensions. IEEE Trans. on Knowl. and Data Eng. 17, 734–749 (2005)
5. Aleman-Meza, B., Nagarajan, M., Ramakrishnan, C., Ding, L., Kolari, P., Sheth, A.P., Arpinar, I.B., Joshi, A., Finin, T.: Semantic analytics on social networks: experiences in addressing the problem of conflict of interest detection. In: Proceedings of the 15th International Conference on World Wide Web, WWW 2006, pp. 407–416. ACM, New York (2006)
6. Cunningham, P., Delany, S.J.: k-nearest neighbour classifiers. Tech. Rep. UCD-CSI-2007-4, School of Computer Science and Informatics, University College Dublin, Ireland (2007)
7. Gauch, S., Madrid, J.M., Induri, S., Ravindran, D., Chadlavada, S.: Keyconcept: A conceptual search engine. Tech. Rep. ITTC-FY2004-TR-8646-37, Information and Telecommunication Technology Center, University of Kansas (2004)
8. Gou, L., Zhang, X.L., Chen, H.H., Kim, J.H., Giles, C.L.: Social network document ranking. In: Proceedings of the 10th Annual Joint Conference on Digital Libraries, JCDL 2010, pp. 313–322. ACM, New York (2010)
9. Kirchhoff, L.: Applying Social Network Analysis to Information Retrieval on the World Wide Web. Ph.D. thesis, the University of St. Gallen, Graduate School of Business Administration, Economics, Law and Social Sciences, HSG (2010)
10. Kirchhoff, L., Stanoevska-Slabeva, K., Nicolai, T., Fleck, M., Stanoevska, K.: Using social network analysis to enhance information retrieval systems. In: Applications of Social Network Analysis (ASNA), Zurich, vol. 7, pp. 1–21 (2008)
11. Matsuo, Y., Mori, J., Hamasaki, M., Nishimura, T., Takeda, H., Hasida, K., Ishizuka, M.: An advanced social network extraction system from the web. Journal of Web Semantics 5(4), 262–278 (2007)
12. Mika, P.: Flink: Semantic web technology for the extraction and analysis of social networks. Journal of Web Semantics 3, 211–223 (2005)
13. Miki, T., Nomura, S., Ishida, T.: Semantic web link analysis to discover social relationships in academic communities. In: Proceedings of the The 2005 Symposium on Applications and the Internet, pp. 38–45. IEEE Computer Society, Washington, DC (2005)
14. Mller-Prothmann, T.: Social network analysis: A practical method to improve knowledge sharing. In: Hands-On Knowledge Co-Creation and Sharing, pp. 219–233 (2007)
15. Tang, J., Zhang, J., Yao, L., Li, J., Zhang, L., Su, Z.: Arnetminer: extraction and mining of academic social networks. In: Proceeding of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD 2008, pp. 990–998. ACM, New York (2008)
16. Yang, Y., Liu, X.: A re-examination of text categorization methods. In: Proceedings of the 22nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR 1999, pp. 42–49. ACM, New York (1999)

A Query Language for WordNet-Like Lexical Databases

Marek Kubis

Uniwersytet im. Adama Mickiewicza,
Faculty of Mathematics and Computer Science,
Department of Computer Linguistics and Artificial Intelligence,
ul. Umultowska 87, 61-614 Poznań, Poland
mkubis@amu.edu.pl

Abstract. WordNet-like lexical databases are used in many natural language processing tasks, such as word sense disambiguation, information extraction and sentiment analysis. The paper discusses the problem of querying such databases. The types of queries specific to WordNet-like databases are analyzed and previous approaches that were undertaken to query wordnets are discussed. A query language which incorporates data types and syntactic constructs based on concepts that form the core of a WordNet-like database (synsets, word senses, semantic relations, etc.) is proposed as a new solution to the problem of querying wordnets.

Keywords: Wordnet, lexical database, domain-specific language.

1 Introduction

Databases that employ data organization inspired by that of WordNet [8] are used in many natural language processing tasks ranging from word sense disambiguation to information extraction and sentiment analysis. The data collected in a wordnet[1] form a network of *synsets*, i.e. semantically related sets of synonymous words. A synset represents a common meaning of the words it contains. Synsets are connected through relations that represent the semantic relationships that hold between the word meanings they represent. Although wordnets and the systems to maintain them have been created for more than 30 natural languages, no common tool has emerged that would allow to access the data collected in the different WordNet-like lexical databases in a uniform manner.

The data collected in a database management system are usually provided to the user through a dedicated programming language (*query language*) that allows the user to create expressions (*queries*) which describe the criteria which the data requested by the user have to fulfill. A wide variety of query languages of both theoretical and practical importance has been proposed for relational,

[1] The terms *wordnet* and *WordNet-like lexical database* are used in the paper interchangeably. The term *WordNet* is used to refer to the database described in [8].

J.-S. Pan, S.-M. Chen, N.T. Nguyen (Eds.): ACIIDS 2012, Part III, LNAI 7198, pp. 436–445, 2012.

object-oriented and semi-structured database management systems. These languages possess data types based on the concepts that appear in the data models for which they have been developed, and provide dedicated syntactic constructs that allow to formulate concise and easy-to-understand queries. We propose a similar approach to data management with respect to WordNet-like lexical databases by introducing WQuery – a query language that incorporates data types and syntactic constructs based on concepts that form the core of a wordnet. The WQuery language interpreter may be freely obtained from the project website located at `http://www.wquery.org`. The interpreter has been used in two projects. First, as a part of an access layer to the lexical database in an NLP system [12]. Second, as a supporting tool in the development of PolNet [19].

2 WordNet-Like Databases

As stated in the introduction, a WordNet-like lexical database is a network of synsets. A synset consists of a set of *words* and a *gloss*. All the words in the synset belong to the same part of speech (POS). Every synset represents a single common meaning of the words it contains. For example, an automobile is represented in WordNet [8] by the set *{ car, auto, automobile, machine, motorcar }* accompanied by the gloss *"a motor vehicle with four wheels; usually propelled by an internal combustion engine (...)"*. A word may occur in several synsets. Every occurrence of the word has an assigned *sense number*. The sense number is a positive integer unique among all the occurrences of the word within the synsets that share the POS. For instance, the word *car* has sense number 1 assigned in the synset above and has sense number 3 in the synset *{ car, gondola }* accompanied by the gloss *"the compartment that is suspended from an airship and that carries personnel and the cargo and the power plant"*. A triple that consists of a word, sense number and POS symbol is called a *word sense*.[2] Synsets are connected through *semantic relations*, i.e. binary relations which represent semantic relationships that hold between the meanings of synsets. For instance, the *hypernymy* relation links synsets representing more general concepts to synsets representing more specific ones. Word senses are connected through *lexical relations*, i.e. binary relations which represent lexical relationships that hold among words. For example, the *antonymy* relation links word senses that represent opposite concepts. Some relations in a wordnet may be inferred from others. For instance, *hyponymy* connects synset A to synset B if and only if *hypernymy* connects synset B to synset A. Besides the data types mentioned above, wordnets also encompass additional data that vary largely among them. For instance, EuroWordNet [20] stores relations that hold between synsets and Inter-Lingual Index, and PolNet [19] provides for synsets examples of usage of the words that belong to them.

[2] In the examples ":" is placed to separate the word, the sense number and the POS symbol of a word sense.

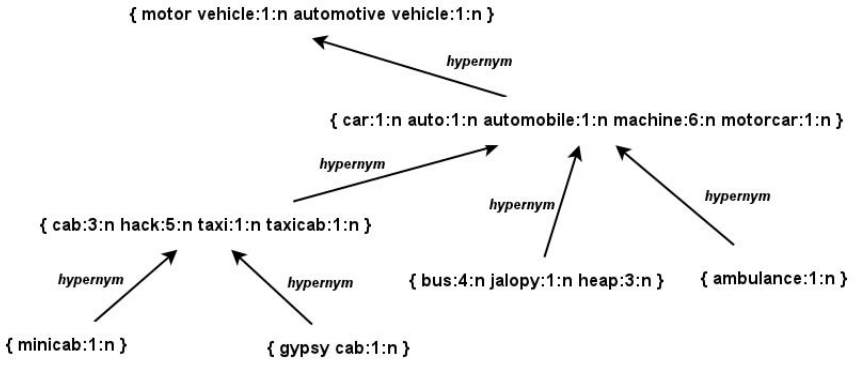


Fig. 1. A fragment of WordNet

3 Wordnet-Specific Queries

Although WordNet-like lexical databases are used in many natural language processing tasks, the class of queries to be answered by a wordnet-specific query language has not been previously investigated. Therefore, we analyzed papers from major wordnet conferences (e.g. [17,16]) and gathered wordnet-specific queries. Since we noticed multiple attempts to use relational databases to store and query wordnets (e.g. [1]), we assumed that the minimal usable wordnet query language should express at least the queries that are representable in relational algebra. Therefore, we present in this section only those queries that require more expressive languages than the relational complete ones.

3.1 Tree Queries

A tree query is a query which for a set of synsets S and a semantic relation r retrieves all paths such that the first synset of a path belongs to S and every other synset of the path is accessible from the previous one through r.

Example 1. (Tree query) Find all paths that link (through hypernymy) synsets containing the word *car* to their hypernyms, hypernyms of their hypernyms, hypernyms of hypernyms of their hypernyms, etc.

Tree queries may be used to build tree views for semantic relations in wordnet browsers and editors. They may also be used to compute structural measures, such as minimum/maximum depth and height of the hypernymy hierarchy[3] and to detect cycles in semantic relations that by definition should be acyclic (e.g. hypernymy). In [12] tree queries are used in the process of building frame-based representations for natural language sentences.

[3] In [6] values of such measures obtained for WordNet are analyzed.

3.2 Reachability Queries

A reachability query is a query which for two given synsets α and β and a set R of semantic relations retrieves all paths that connect the synset α to the synset β through relations from the set R.

Example 2. (Reachability query) Find all paths that connect through hypernymy the synset that contains the fourth noun sense of the word *bus* to the synset that contains the first noun sense of the word *vehicle*.

The results of reachability queries are used to determine values of path-based similarity and relatedness measures, which are further exploited in the word sense disambiguation task [14]. For instance, the Leacock-Chodorow measure [13] of similarity between words a and b is defined by the formula $sim_{ab} = max[-log(Np/2D)]$, where Np means the number of nodes of a path p between a and b, and D means the depth of hyponymy. The results of reachability queries are also determined in WordNet-based anaphora resolution methods [7]. As in the case of tree queries, one may formulate reachability queries that detect cycles in semantic relations.

3.3 Least Common Subsumers

The least common subsumer of two synsets α and β is a common hypernym γ of α and β such that no other hypernym of α and β is placed lower in the hypernymy hierarchy than γ.

Example 3. (Least Common Subsumer) The synset *{ car:1:n auto:1:n automobile:1:n auto:1:n automobile:1:n }* presented in Figure 2 is the least common subsumer of *{ bus:4:n jalopy:1:n heap:3:n }* and *{ minicab:1:n }*.

As in the case of reachability, the notion of the least common subsumer is exploited in similarity and relatedness measures (e.g. [14] the Resnik, Jiang-Conrath and Lin measures).

4 Related Work

A tool called Hydra that provides a modal logic based language for querying wordnets has been described in [15]. The Hydra language may be translated to relational calculus by adjusting the standard translation from the modal language to the first-order language defined in [3]. Therefore, one cannot express in Hydra any of the wordnet-specific queries described in Section 3 because these queries involve computation of the transitive closures of relations, which are not computable in relational calculus. The *WN* language [11] is another modal logic based formalism designed to query WordNet-like databases. *WN* represents the transitive closure of hypernymy as a separate binary relation. However, it does not provide operators that would allow to compute the entire paths that transitively connect synsets through arbitrary chosen semantic relations. Hence, one

cannot formulate in *WN* queries defined in Section 3.2. Neither *WN* nor Hydra encompass arithmetic expressions. Therefore, they are unable to answer aggregate queries such as "How many synsets exist in the wordnet?" or "What is the average polysemy?", and they cannot be used to compute the Leacock-Chodorow measure mentioned in Section 3.2. DEBVisDic [10] – a tool for browsing and editing wordnets – provides an option to search for synsets by specifying the word form, the word sense and several other properties that conform to the underlying XML representation of a wordnet. These search criteria may be combined together by using logic operators. Such queries may be expressed in WQuery using the synset generator followed by a filter with a suitable condition (Section 6.2). The wordnet-specific queries defined in Section 3.2 and 3.3 are not directly expressible using the search form of DEBVisDic.

As for general purpose tools adapted to query wordnets, relational databases are reported to have been used in several projects, e.g. [1]. Since the queries described in Section 3 involve computation of transitive closures, they cannot be formulated in relational complete languages. Besides the use of relational databases, there have been attempts to store wordnets in XML [10,18] and RDF [9]. Thus, one may consider using one of the languages designed for these models, such as XQuery or SPARQL. However, the XML and RDF query languages do not include the wordnet-specific data types described in Section 2. Another option is to access a wordnet from a general purpose programming language through one of the APIs mentioned in [1] or the DEBVisDic server API. However, regardless of the choice, general purpose tools cannot be more expressive than WQuery since it expresses all computable queries over the wordnet data model, as stated in Section 6.4.

Path expressions, which form the core of the WQuery language, are similar to constructs available in other regular path languages, such as Lorel [2] and XPath [5], but the preservation of variable bindings in functions is not, to the best of our knowledge, supported by any of the tools mentioned above.

5 Data Model

According to the description of a WordNet-like lexical database presented in Section 2, a wordnet query language has to operate on such domain-specific data types as synsets, word senses, words, glosses, POS symbols and sense numbers. One may notice that the relationships which define the structure of a wordnet may be modeled as a set of binary relations between the aforementioned data types that consists of:

1. The relation *synset* that links word senses to their synsets.
2. The relation *word* that links word senses to their words.
3. The relation *sensenum* that links word senses to their sense numbers.
4. The relation *pos* that links word senses to their part of speech symbols.
5. The relation *gloss* that links synsets to their glosses.

However, in order to embrace extensions specific to particular existing wordnets, we amplify this model in several ways. First, we include two additional data types

– floating point numbers and booleans. Secondly, we assume that words, glosses and POS symbols are encompassed by the generic character string data type and that sense numbers belong to the integer data type. Thirdly, we represent the inferred relations by relation names bound to regular expressions composed over other previously defined relations.

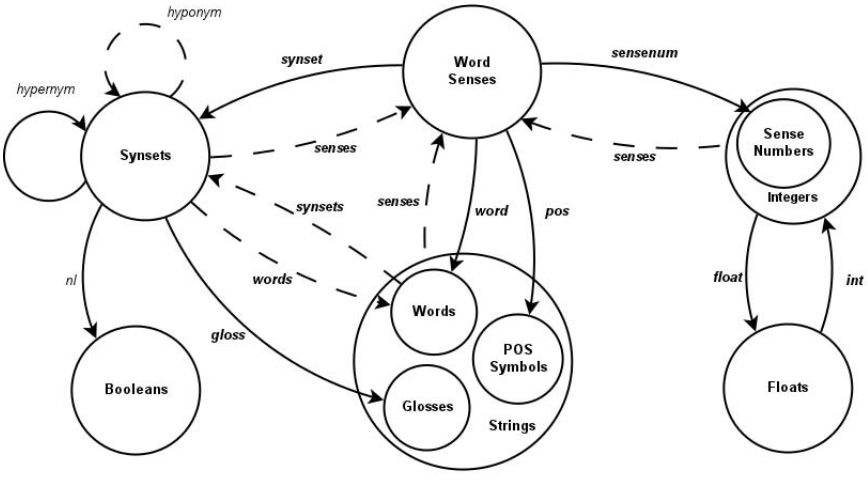


Fig. 2. An instance of the WQuery wordnet model. The mandatory relations are marked by boldfaced labels. The inferred relations are represented by dashed edges.

6 Syntax and Semantics

6.1 Path Expressions

The syntax and semantics of the WQuery language are focused on searching for paths in instances of the data model defined in the previous section. Paths are represented at the syntactic level by *path expressions*, and at the semantic level by tuples that consist of values of data types defined in Section 5, interleaved with the names of the relations that connect the consecutive values. A path expression consists of one or more *steps*, such that the first one (*generator*) describes a set of nodes and the following ones (*transformations*) extend the paths defined by the preceding expressions by attaching edges and nodes to them. For example, the expression

```
{car}.(hypernym|meronym).words
```

consists of a generator `{car}` which represents the set of synsets that contain the word *car*, the transformation `.(hypernym|meronym)` which extends one-element paths generated by `{car}` with edges reachable through either the `hypernym` or `meronym` relation, and the transformation `words` which extends the resulting

paths with words that belong to the synsets reached by the previous transformation. The following tuples belong to the result of the query formulated above[4]

```
{ car:1:n ... } meronym { air bag:1:n } words 'air bag'
{ car:1:n ... } meronym { bumper:2:n } words bumper
{ car:3:n ... } hypernym { compartment:2:n } words compartment
```

Transformations may involve conventional regular operators, such as `*` (zero or more), `+` (one or more), `|` (alternative) and parentheses. One may also use the `^` operator to traverse relations in the opposite direction.

6.2 Variable Bindings

In order to select data from paths, one may augment a path expression with variables placed after a step. For instance, the path expression

```
{car}$a.hypernym+@P$b
```

will bind the variable `$b` to the last transitive hypernym on the path, the path variable `@P` to all elements of the path generated by the transformation `.hypernym+` that precede `$b` and the variable `$a` to a node generated by `{car}`.

Values bound to variables may be referenced in *filters* and *projections* – two constructs that can be placed after a step of a path expression in order to modify the result. A filter consists of a logical condition enclosed in square brackets. If the condition inside the filter holds for a path, then the path is added to the resulting multiset, otherwise it is eliminated. For example, the query

```
{car}.hypernym+$a[$a = {vehicle:1:n}]
```

finds paths that connect synsets that contain the word *car* to the synset that contains the word sense *vehicle:1:n* through the `hypernym` relation. Filters that compare a value bound to the last step of the path with a value of a particular generator may also be abbreviated by placing the generator after the dot instead of the bracketed expression. Thus, the query above may be reformulated as

```
{car}.hypernym+.{vehicle:1:n}
```

A projection consists of a path expression enclosed in `<` and `>`. The result of the evaluation of the enclosed expression replaces the path for which the projection is evaluated. For example, the query

```
{car}$a.hypernym.meronym$b<$a,$b>
```

finds all pairs that consist of a synset that contains the word *car* and a meronym of its hypernym.

[4] The queries in the paper are invoked against WordNet [8], version 3.0.

6.3 Functions

WQuery provides a set of built-in functions which take as arguments and return as values arbitrary multisets of tuples. Among them there are conventional aggregate operators, such as `count`, `max` and `avg`, mathematical functions, such as `log` and `exp`, and special path-oriented operators, such as `shortest` and `longest`, which return the shortest and longest tuple of a multiset, respectively. What differs WQuery functions from their counterparts in other query languages is that the functions that do not modify, remove or replace elements of their arguments preserve variable bindings. Thus, the query

```
longest({car:5:n}.hypernym+$a)
```

not only returns the longest path from `{car:5:n}` to the top of the hypernymy hierarchy, but also binds the synset on the top to the variable `$a`.

6.4 Other Constructs

Although the queries presented in Section 3 may be formulated using the constructs presented in the previous sections, we have introduced additional constructs to make the language robust. First, we extended the language with multipath operators that are counterparts of relational algebra operators. The `union`, `intersect`, `except` and cross product (`,`) operators treat paths as tuples and make it possible to translate queries formulated in relational complete languages to WQuery. Secondly, we added conventional arithmetic operators to make it easier to compute similarity measures mentioned in Section 3 directly in WQuery. Thirdly, we introduced imperative constructs such as `if-else` statements, `while` loops and sequential execution blocks that made it possible to prove that the WQuery language can express any query computable over an instance of the wordnet data model. Due to the limited space, we do not present the formal proof here, but it can be easily derived by constructing the translation from the QL language [4] to WQuery.

7 Wordnet-Specific Queries in WQuery

We will now show how the wordnet-specific queries described in Section 3 may be expressed in WQuery.

7.1 Tree Queries

Let `g` be a generator that represents a set of synsets S and `r` be the name of a semantic relation r. The tree query of S under r is defined by the expression

```
g.r+
```

For instance, to answer the example query presented in Section 3.1 one may formulate the following expression

```
{car}.hypernym+
```

7.2 Reachability Queries

Let `g` and `h` be generators for synsets α and β, respectively. Let `r_1`, ..., `r_k` be the names of the semantic relations in set R. The reachability query from α to β with respect to R is fulfilled by the expression

```
g.(r_1|...|r_k)*.h
```

For example, to find paths that connect the synset that contains the word sense *bus:4:n* with the synset that contains the word sense *vehicle:1:n* through hypernymy one may formulate the expression

```
{bus:4:n}.hypernym*.{vehicle:1:n}
```

7.3 Least Common Subsumers

Let `g` and `h` be generators for synsets α and β, respectively. The least common subsumer of α and β is determined by the query

```
longest((g.hypernym*$a<$a>
    intersect h.hypernym*$b<$b>)$s.hypernym*)<$s>
```

For instance, the least common subsumer of the synsets that contain the senses *minicab:1:n* and *bus:4:n* is computed by the query

```
longest(({minicab:1:n}.hypernym*$a<$a>
    intersect {bus:4:n}.hypernym*$b<$b>)$s.hypernym*)<$s>
```

8 Conclusion

We have presented in the paper several types of wordnet-specific queries that arise in natural language processing tasks, such as word sense disambiguation and relatedness measurement. We found that none of the existing wordnet-specific tools are capable of expressing all of the analyzed queries. Therefore, we proposed a new solution to the problem of querying WordNet-like lexical databases. The solution is based on the data model that incorporates wordnet-specific data types and a query language that utilizes the concept of a path. We have shown that by using the WQuery language one can formulate all wordnet-specific queries described in Section 3. In the future we would like to introduce optimization techniques for query evaluation based on the structure of a WordNet-like database.

Acknowledgments. The author is a scholarship holder within the project "Scholarship support for Ph.D. students specializing in majors strategic for Wielkopolska's development", Sub-measure 8.2.2 Human Capital Operational Programme, co-financed by European Union under the European Social Fund.

References

1. WordNet - Related Projects, http://wordnet.princeton.edu/wordnet/related-projects/ (access date: April 28, 2011)
2. Abiteboul, S., Quass, D., McHugh, J., Widom, J., Wiener, J.L.: The Lorel Query Language for Semistructured Data. International Journal on Digital Libraries 1(1), 68–88 (1997)
3. Blackburn, P., van Benthem, J.: Modal Logic: a Semantic Perspective. In: Blackburn, P., et al. (eds.) Handbook of Modal Logic, pp. 1–84. Elsevier (2007)
4. Chandra, A.K., Harel, D.: Computable Queries for Relational Data Bases. Journal of Computer and System Sciences 21(2), 156–178 (1980)
5. Clark, J., DeRose, S.: XML path language (XPath) version 1.0. W3C recommendation, W3C (1999), http://www.w3.org/TR/1999/REC-xpath-19991116/
6. Devitt, A., Vogel, C.: The Topology of WordNet: Some Metrics. In: Sojka, et al. (eds.) [17], pp. 106–111
7. Fan, J., Barker, K., Porter, B.: Indirect Anaphora Resolution as Semantic Path Search. In: K-CAP 2005: Proceedings of the 3rd International Conference on Knowledge Capture, pp. 153–160. ACM Press (2005)
8. Fellbaum, C. (ed.): WordNet: An Electronic Lexical Database. MIT Press, Cambridge (1998)
9. Graves, A., Gutierrez, C.: Data Representations for WordNet: A Case for RDF. In: Sojka, et al. (eds.) [16], pp. 165–169
10. Horak, A., Pala, K., Rambousek, A., Povolny, M.: DEBVisDic - First Version of New Client-Server Wordnet Browsing and Editing Tool. In: Sojka, et al. (eds.) [16]
11. Koeva, S., Mihov, S., Tinchev, T.: Bulgarian Wordnet – Structure and Validation. Romanian Journal of Information Science and Technology 7(1-2), 61–78 (2004)
12. Kubis, M.: An Access Layer to PolNet – Polish WordNet. In: Vetulani, Z. (ed.) LTC 2009. LNCS, vol. 6562, pp. 444–455. Springer, Heidelberg (2011)
13. Leacock, C., Chodorow, M.: Combining Local Context and WordNet Similarity for Word Sense Identification. In: Fellbaum (ed.) [8], ch. 11, pp. 265–283 (1998)
14. Patwardhan, S., Banerjee, S., Pedersen, T.: Using Measures of Semantic Relatedness for Word Sense Disambiguation. In: Gelbukh, A. (ed.) CICLing 2003. LNCS, vol. 2588, pp. 241–257. Springer, Heidelberg (2003)
15. Rizov, B.: Hydra: a Modal Logic Tool for Wordnet Development, Validation and Exploration. In: Calzolari, N., et al. (eds.) Proceedings of the Sixth International Conference on Language Resources and Evaluation, LREC 2008 (2008)
16. Sojka, P., et al. (eds.): Proceedings of the Third International WordNet Conference – GWC 2006. Masaryk University, Brno (2005)
17. Sojka, P., et al. (eds.): Proceedings of the Second International WordNet Conference-GWC 2004. Masaryk University, Brno (2003)
18. Soria, C., Monachini, M., Vossen, P.: Wordnet-LMF: Fleshing out a Standardized Format for Wordnet Interoperability. In: Proceeding of the 2009 International Workshop on Intercultural Collaboration, pp. 139–146. ACM, New York (2009)
19. Vetulani, Z., Kubis, M., Obrębski, T.: PolNet - Polish WordNet: Data and Tools. In: Calzolari, N., et al. (eds.) Proceedings of the Seventh International Conference on Language Resources and Evaluation. ELRA, Valletta (2010)
20. Vossen, P.: EuroWordNet: general document. Vrije Universiteit, Amsterdam (2002), http://dare.ubvu.vu.nl/handle/1871/11116

Reversible Data Hiding with Hierarchical Relationships

Hsiang-Cheh Huang[1], Kai-Yi Huang[1], and Feng-Cheng Chang[2]

[1] National University of Kaohsiung, 700 University Road,
Kaohsiung 811, Taiwan, R.O.C.
hch.nuk@gmail.com
[2] Tamkang University, 180 Linwei Road, Jiaosi, Ilan 262, Taiwan, R.O.C.
135170@mail.tku.edu.tw

Abstract. In this paper, we propose a new method for reversible data hiding by employing the hierarchical relationships of original images. Considering the ease of implementation and the little overhead needed for decoding, we employ the histogram-based scheme with extensions for reversible data hiding. By utilizing the hierarchical structure and corresponding histograms of difference values, global and local characteristics of original images can be utilized for hiding more capacity with acceptable quality of output image. With our method, better performances can be obtained in enhanced image quality, embedding capacity, and comparable amount of side information. More importantly, the reversibility of our method is guaranteed, meaning that original image and hidden message can both be perfectly recovered at the decoder. Simulation results demonstrate that proposed method in this paper outperforms those in conventional histogram-based algorithms.

Keywords: Reversible data hiding, hierarchical structure, quad, image quality, capacity.

1 Introduction

Watermarking researches has emerged for around 15 years, and reversible data hiding is a recently developed branch in watermarking researches. On the one hand, for conventional watermarking, at the encoder, the secret data should be embedded into the original multimedia contents, digital images in most cases, by use of algorithms developed by researchers, and then the watermarked media can be transmitted to the receiver. Data loss or intentional attacks may be experienced during transmission. After reception of the delivered watermarked media, only the secret data need to be extracted [1]. On the other hand, for reversible data hiding, data embedding is similar to its counterpart with conventional watermarking applications. Different from conventional watermarking applications, for reversible data hiding, after the reception of marked media, both the original content and the embedded secret data need to be recovered and extracted perfectly with a reasonable amount of side information [2, 3, 4]. And this is the origin of the term “reversible” comes from. Due to this kind of characteristics, during the transmission, the watermarked media need to be kept intact.

J.-S. Pan, S.-M. Chen, N.T. Nguyen (Eds.): ACIIDS 2012, Part III, LNAI 7198, pp. 446–455, 2012.

In this paper, we focus on enhancing the histogram-based reversible data hiding algorithm [2] with hierarchical structures. Reversible data hiding methods, which will be described in Sec. 2, have their limitations and drawbacks. More importantly, few methods take the characteristics of original images into account. Here we make use of the hierarchical structure of original image for obtaining the larger number of secret bits for embedding, with similar quality of the output images. Simulation results reveal that the algorithm proposed in this paper outperforms conventional one by use of test images.

This paper is organized as follows. In Sec. 2, we describe fundamental concepts of reversible data hiding algorithms, including the histogram-based and difference-expansion-based ones, and the requirements for algorithm design. Then, in Sec. 3, we present the way to make use of the difference values for making reversible data hiding with the histogram-based algorithm. Proposed algorithm by utilizing the hierarchical structure for reversible data hiding is depicted in Sec. 4. Simulation results are demonstrated in Sec. 5, which point out the guaranteed image quality, the more embedding capacity, and the less side information needed for the proposed algorithm. Finally, conclusion of this paper is addressed in Sec. 6.

2 Relating Methods

Practical schemes for making reversible data hiding possible can be categorized into two branches, one is by modifying the histogram of original image, and the other is to intentionally adjust the difference value between neighboring pixel pairs. To assess the performances of these schemes, one can observe the output image quality (called imperceptibility), the number of bits for embedding (called capacity), and the side information necessary for decoding, to determine how good the schemes are. Brief descriptions of the two branches and discussions about performance assessments and corresponding limitations are stated as follows.

2.1 Histogram-Based Reversible Data Hiding

Histogram-based reversible data hiding is famous for its ease of implementation [2]. By intentionally modifying the histogram of original image, secret data can reversibly be hidden. For obtaining the largest capacity, the luminance of the peak value in the histogram is selected, and such a value, one byte in length, would be served as the side information for decoding. By moving the portion larger to the luminance of the peak to the right by one, the secret data can be embedded accordingly.

We can observe that only the moving of certain portion in the histogram is required, and no calculation is needed. Moreover, embedding of secret data causes the mean square error (MSE) of one in the worst case, leading to the guaranteed output image quality of at least $10 \cdot \log\left(\frac{255^2}{1}\right) = 48.13$ dB.

On the contrary, the number of bits for embedding, or the capacity, might not be enough for residing all the data to be hidden. Capacity is limited by the number of occurrences of the max point.

2.2 Difference-Expansion-Based Reversible Data Hiding

The difference expansion (DE) method is one of the earliest schemes for reversible data hiding [3, 4]. It follows the concepts directly from wavelet transforms by turning the spatial pixel values into frequency coefficients. In DE, every two neighboring pixels should be grouped together as a pair. There are two principles for reversible data hiding with DE. The first one is keeping the same average value of the pair before and after data embedding. And the second one is to multiply the luminance difference value of the pair by two, and add the secret bit into the multiplied difference. Thus, we can easily see the origin of the term of "difference expansion".

Considering the representation of images, the luminance value of each pixel should lie between 0 and 255. Under normal conditions when the luminance values of output pixel pair lie between 0 and 255, one bit can be embedded, leading to the capacity of 0.5 bit per pixel (bpp). On the contrary, if the luminance values of output pixels lie outside 0 and 255, such positions, called the location map that are regarded as the side information, need to be recorded in order to keep the reversibility of the algorithm.

Therefore, we can easily observe that with DE, calculations including addition and multiplication are required for data embedding. Also, the side information (or the location map) reduces the effective capacity for data embedding.

Due to the simplicity for implementation, we concentrate on the modification of the former scheme in this paper, while keeping the capacity of DE, or 0.5 bpp, as another goal for the development of algorithm.

3 Data Hiding with Difference of Histogram

Besides altering the histogram of original image directly, we consider the local characteristics by using the quad modification, and the global characteristics with the histogram alteration method. Another advantage of our algorithm is that we take the advantages of the ease of implementation by utilizing the histogram-based reversible data hiding algorithm. We briefly describe existing scheme proposed by our research group in [5] to serve as the baseline for the proposed algorithm in this paper.

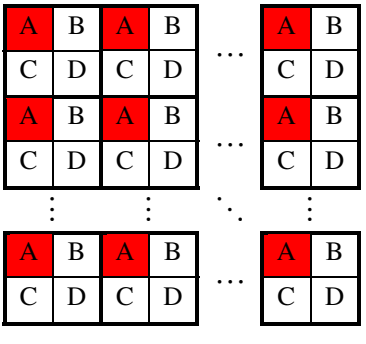


Fig. 1. The quad structure

We can classify the original image into small blocks for making reversible data hiding possible. A quad is a 2×2 block, in which four pixels are placed at regular positions in 'A', 'B', 'C', and 'D' in Fig. 1. Each position in Fig. 1 represents one pixel. For the original image with the size of $M\times N$, it will compose of $\frac{M}{2}\times\frac{N}{2}$ quads.

By following the concept of difference expansion of quads (DEQ) in [3], we first keep the pixel values at positions 'A' the same, and calculate the other three difference values in the quad by:

$$d_i = \mathrm{lum}(i) - \mathrm{lum}(A); \quad i = B, C, D. \tag{1}$$

Here, the luminance value of corresponding position is denoted by $\mathrm{lum}(\bullet)$. Next, the three difference values of every quad in the original image are recorded, and it makes a total of $\frac{M}{2}\times\frac{N}{2}\times3$ difference values. The histogram of all the $\frac{M}{2}\times\frac{N}{2}\times3$ difference values is prepared, and data embedding with the histogram-based scheme can be utilized. With this setting, three bits can be embedded into four pixels, leading to the capacity of 0.75 bpp.

For natural images, luminance values of neighboring pixels tend to be similar, and then the difference values in Eq. (1) are likely to be concentrated around zero. With the descriptions in Sec. 2.1, the maximal value in histogram directly influences the embedding capacity. Hence, the larger capacity by utilizing the difference is expected.

Next, the pre-determined threshold value δ is selected for making data embedding possible, and it also serves as the side information at the decoder. The threshold value should be a positive integer. Data embedding should meet one of the following cases.

Case 1. If $d_B, d_C, d_D > \delta$,

$$d_i' = d_i + 1; \quad i = B, C, D. \tag{2}$$

Case 2. If $d_B, d_C, d_D \le -\delta$,

$$d_i' = d_i - 1; \quad i = B, C, D. \tag{3}$$

Case 3. If $-\delta+1 \le d_B, d_C, d_D \le \delta$, the values are kept the same. That is,

$$d_i' = d_i; \quad i = B, C, D. \tag{4}$$

From Case 1 to Case 3, histogram occurrences at difference values of $(\delta+1)$ and $-\delta$ are intentionally set to zero. Embedding meets one of following conditions.

E1. For embedding bit '1',

If $d_B, d_C, d_D = \delta$,

$$d_i'' = d_i' + 1; \quad i = B, C, D. \tag{5}$$

If $d_B, d_C, d_D = -\delta+1$,

$$d_i'' = d_i' - 1; \quad i = B, C, D. \tag{6}$$

E2. For embedding bit '0', keep the difference values the same. That is,

$$d_i'' = d_i'; \qquad i = B, C, D. \tag{7}$$

Finally, the output image should be obtained by

$$\mathrm{lum}(i') = d_i'' + \mathrm{lum}(A); \quad i = B, C, D. \tag{8}$$

We observe that the value of δ plays the role of the secret key in reversible data hiding with only a few bits of overhead. There is one drawback for the proposed algorithm. Under the extreme cases when difference values equal to -255 or 255, the overflow problem would occur. To prevent this from happening, location of quad for such a situation should be recorded, similar to the role of location map in DE scheme.

On the contrary, data extraction is the reverse process to the data embedding procedure. They can be briefly described by the following steps.

Step 1. In the received image, calculate the three difference values in each quad.

Step 2. Check the difference values.

- If the difference value equals to $(\delta+1)$ or $-\delta$, output bit '1' as the secret data, and set the difference value to δ or $(-\delta+1)$, respectively.
- If the difference value equals δ or $(-\delta+1)$, output bit '0' as the secret data, and keep the difference value unchanged.
- For all other values, keep them unchanged.

Step 3. Generate the original difference histogram.

- If difference values are larger than δ, decrease the value by one.
- If difference values are smaller than $(-\delta+1)$, increase the value by one.
- For all the other difference values, keep them unchanged.

Step 4. Recover the original by adding the difference value back to $\mathrm{lum}(A)$.

From the descriptions of data embedding and extraction procedures above, we can find that both the difference value and the histogram modification schemes are all considered. We can take the advantages from the two major branches in reversible data hiding.

4 Proposed Schemes

Here we employ the hierarchical relationships of original image to make reversible data hiding possible. By carefully modifying the histogram, better performances can be obtained under the reversibility criterion.

The fundamental concept for reversible data hiding is to elaborately hide the secret information into the original image. We take the random bitstreams with equal probabilities of bit-0 and bit-1 to serve as the information to be hidden.

Here we present the reversible data embedding scheme. We consider the hierarchical structure first. Like the classification of quad in Sec. 3, we split the original image into quads. As the extension to Fig. 1, according to the locations of the 2×2 blocks, one reference pixel is selected, and three difference values in one block can be calculated. Position 'A' in Fig. 1 would be changed to the red pixels according to the positions of the quad. This structure serves as the first layer for data embedding. If the quad is in odd row and odd column, position 'D' serves as the reference, shown in Fig. 2(a). Similar manners can be performed for Figs. 2(b), 2(c), and 2(d), respectively. By gathering all the quads in the original image altogether, as depicted in Fig. 3, hierarchical structure can be displayed.

Next, data embedding can be performed subsequently with Eq. (1) to Eq. (8). We denote δ_n as the threshold for the n^{th} layer. Demonstration of $\delta_1 = 1$ for data embedding can be found in Fig. 4.

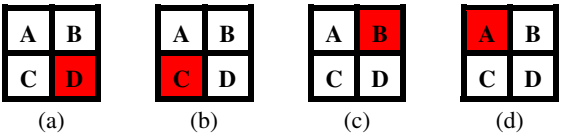


Fig. 2. Reference locations in red for the first-layer cell, or the quad. (a) Odd row, odd column. (b) Odd row, even column. (c) Even row, odd column. (d) Even row, even column.

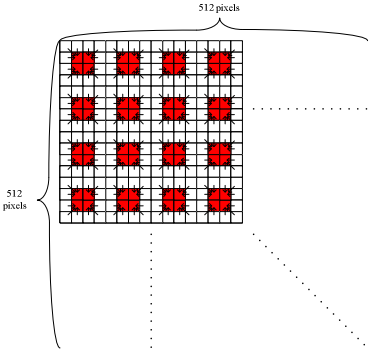


Fig. 3. Illustration of the first-layer cell of the whole image by gathering Fig. 2 altogether

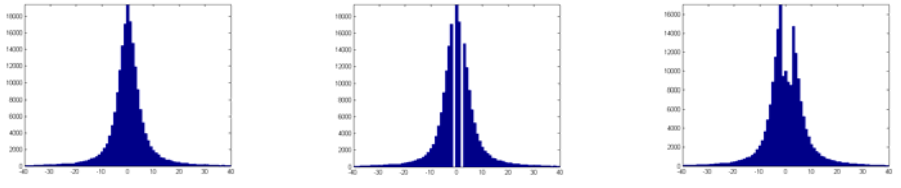


Fig. 4. Demonstration of data embedding for $\delta_1 = 1$ in the first layer. (a) Histogram of difference values. (b) Empty the difference values at 2 and -1. (c) Embed secret data.

We expect that performances by using the first layer between proposed algorithm and [5] would be similar, because the pixels in red in Fig. 2 lie on the same or adjacent positions to their counterparts in Fig. 1. We are going to arrange the data embedding in the second layer. As depicted in Fig. 5, the width and height of quad in the second layer have become doubled, meaning that the 4×4 block serves as one cell for data embedding. Following similar steps with Fig. 2, the pixel positions in red in Fig. 5 are served as the reference, and the differences of luminance values between the three white positions and the reference position are calculated. For keeping the output image quality, luminance values of remaining 12 positions, denoted in grey in Fig. 5, are kept intact.

By following the same concept, the layers for data embedding can further be explored. For instance, the cell for embedding in the third layer is 8×8, only the four pixels in the central are chosen for data embedding, and the other 60 are left intact.

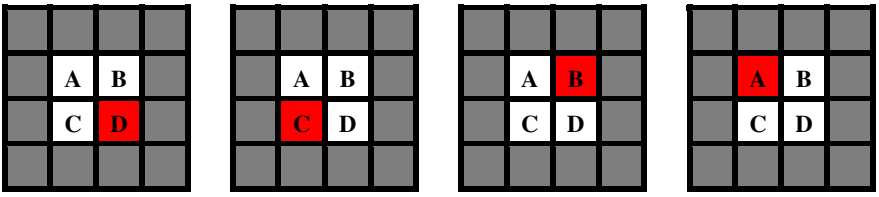


Fig. 5. Reference locations in red for the second-layer cell, or 4×4 blocks. (a) Odd row, odd column. (b) Odd row, even column. (c) Even row, odd column. (d) Even row, even column.

With this manner, we observe that the capacity decreases with the increase of the layer number. Thus, besides looking for the maximally allowable capacity, we can adaptively choose the specific layers for data embedding to look for the balance between output image quality and embedding capacity.

Data extraction and original recovery are the reverse procedures to the embedding counterpart. The threshold values for each layer, δ_i, $\forall i$, are delivered to the decoder as the side information. Next, the four steps for the extraction of secret data and the recovery of original image in Sec. 3 can be performed sequentially. Simulation results will demonstrate the effectiveness of the proposed algorithm.

5 Simulation Results

In our simulations, we use the Lena test image, with the size of 512×512 in Fig. 6(a), for both subjective and objective evaluations. For achieving the more capacity than DE (or 0.5 bpp at most) and DEQ (or 0.75 bpp at most), we choose four layers for embedding secret data, and the thresholds of each layer range are $1\le\delta_i\le13$, $i=1, 2, 3, 4$. With the proposed algorithm, by using the hierarchical structure, 206641 bits, or 0.7883 bpp of capacity can be reached, with the PSNR value of 32.392 dB. Fig. 6(b) is the output image for subjective evaluation. We can also make comparisons between the histogram of original image in Fig. 6(c), and that of the output image in Fig. 6(d). We can easily see that the histogram in Fig. 6(d) has been modified from Fig. 6(c).

Results with the use of one layer only are depicted in Table 1. We can easily observe that the number of cells in one layer is four times to that of the next layer. Thus, the fewer capacity in higher layer can be expected. Also, because fewer bits is embedded into higher layers, corresponding PSNR values become much higher.

Next, we can compare the histograms of luminance difference in the four layers from Fig. 7(a) to Fig. 7(d), respectively. Even though the histograms look similar, we need to note that the peak values of the four figures, corresponding to each layer, are quite different. In comparison with Table 1, we can adaptively choose the required capacity based on the capacity and threshold in each layer, and then both the capacity and the output image quality can both be balanced.

Table 1. Results with the use of one layer only

Hierarchy	δ value	PSNR (dB)	Capacity (bits)
The first layer	1	50.307	34614
	2	51.276	29448
	3	52.280	23530
	4	53.280	18108
The second layer	1	56.126	7305
	2	56.902	6430
	3	57.712	5308
	4	58.485	4198
The third layer	1	61.945	1327
	2	62.509	1282
	3	63.091	1133
	4	63.657	943
The fourth layer	1	67.828	272
	2	68.270	265
	3	68.691	226
	4	69.096	205

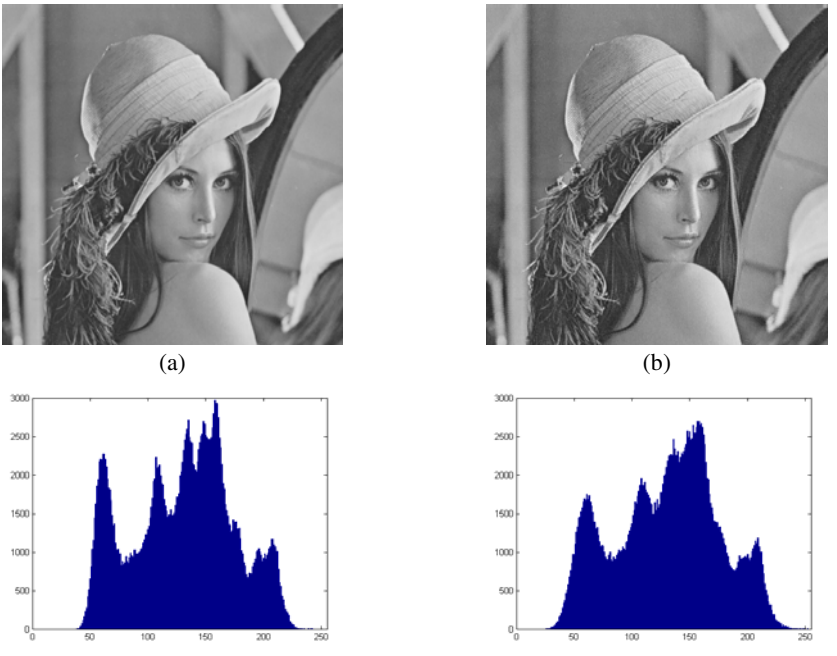
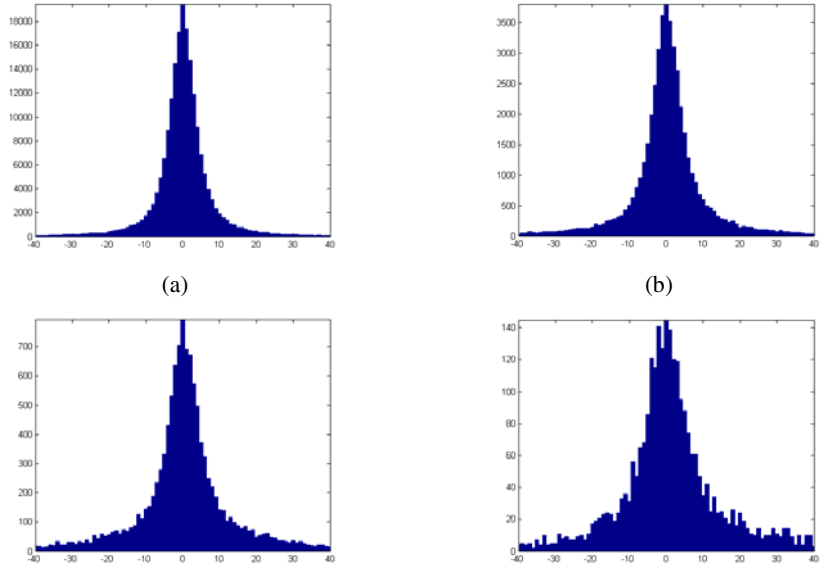


Fig. 6. Simulation results for Lena. (a) Original image. (b) Output image, PSNR = 32.392 dB after embedding the maximal amount of 206641 bits, or 0.7883 bpp. (c) The histogram of original image. (d) The histogram of output image.

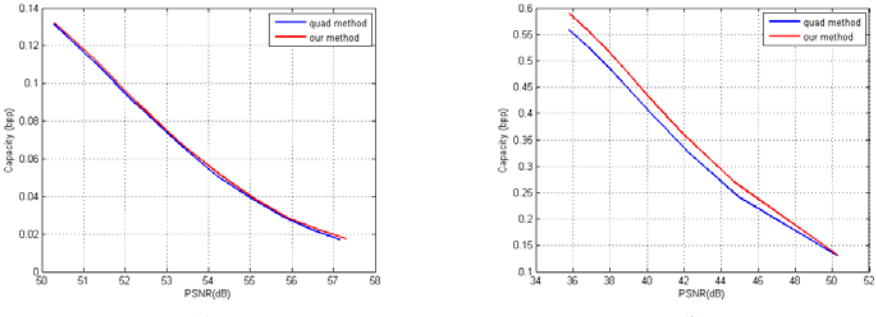


(a) (b)

(c) (d)

Fig. 7. Histogram of difference values in different layers after data embedding. Note the difference of peak value of each layer. (a) The first layer. (b) The second layer. (c) The third layer. (d) The fourth layer.

Finally, we can compare the performances between the proposed method and conventional method in [5]. Figure 8(a) illustrates the relationships between capacity and output image quality by using one layer only. The two curves almost lap together, because the reference positions in Fig. 1 and Fig. 2 are very near or identical. If we use four layers altogether for data hiding, our method outperforms that of conventional method. However, due to the fact that more layers can be employed for data embedding, and one layer corresponds to one threshold value, by use of our algorithm, the slight increase in side information would be necessary for the decoder.



(a) (b)

Fig. 8. Performance comparisons. (a) With one layer. (b) With four layers.

6 Conclusions

In this paper, we proposed an effective algorithm for reversible data hiding by use of the hierarchical structure of original images. At the encoder, by utilizing the difference value of the histogram, secret data can be effectively embedded into the original image. At the decoder, with the provision of threshold value, reversibility of the proposed algorithm can be verified. Considering the inherent characteristics of original images, more capacity can be embedded with the same amount of output image quality in PSNR vales. In addition, the amount of side information needed for the recovery of original image and hidden data is comparable to that with the conventional schemes. Finally, how to systematically and effectively choose the number of layers for data embedding, the threshold value of each layer, and the balance between output image quality and the embedding capacity will be further explored in the future.

Acknowledgements. This work is supported in part by the National Science Council of Taiwan, R.O.C., under grants NSC100-2220-E-390-002 and NSC100-2221-E-390-013.

References

1. Huang, H.C., Fang, W.C.: Metadata-Based Image Watermarking for Copyright Protection. Simulation Modelling Practice and Theory 18(4), 436–445 (2010)
2. Ni, Z., Shi, Y.-Q., Ansari, N., Su, W.: Reversible data hiding. IEEE Trans. Circuits Syst. Video Technol. 16(3), 354–362 (2006)
3. Tian, J.: Reversible data embedding using a difference expansion. IEEE Trans. Circuits Syst. Video Technol. 13(8), 890–896 (2003)
4. Alattar, A.M.: Reversible watermark using the difference expansion of a generalized integer transform. IEEE Trans. Image Process. 13(8), 1147–1156 (2004)
5. Huang, H.C., Chen, T.W., Chang, F.C.: Adjacent quad modification algorithm for reversible data hiding. In: Proc. Int'l. Conf. on Intelligent Information Hiding and Multimedia Signal Processing, Darmstadt, Germany, pp. 171–174 (2010)

Effect of Density of Measurement Points Collected from a Multibeam Echosounder on the Accuracy of a Digital Terrain Model

Wojciech Maleika, Michal Palczynski, and Dariusz Frejlichowski

West Pomeranian University of Technology, Szczecin
Faculty of Computer Science and Information Systems,
Zolnierska 52, 71-719, Szczecin, Poland
{wmaleika,mpalczynski,dfrejlichowski}@wi.zut.edu.pl

Abstract. Digital terrain models (DTMs), finding a wide range of applications in the exploration of water areas, are mainly created on the basis of bathymetric data from a multibeam echosounder. The estimation of DTM accuracy dependent on the choice of the survey parameters is difficult due to the lack of reference surface. These authors have developed the methodology of simulation called *virtual survey*, which enables examining how various parameters of the echosounder, survey and DTM construction algorithms affect the errors of the created models. They are aimed at precise estimation of the model accuracy and the optimization of depth measurement work. The article includes the results of the examination of the effect of parameters determining the density of measurement points on the accuracy of the obtained GRID model. It has been proved that a significant reduction of recorded data density leads to only a slight increase in the modeling error, which makes the bathymetric survey much more cost-effective.

Keywords: Digital terrain model, bathymetric survey, multibeam echosounder.

1 Introduction

Detailed depth data are usually needed for the exploitation of submarine resources. The relevant information, often visualized and processed with computer software, enables comprehensive in-depth analyses. Contrary to land areas, where surveying methods or global positioning systems allow to obtain a highly accurate height of any point, depth measurements continue to be inaccurate and rather expensive, and in many water areas quickly become outdated due to changes in the seabed relief. Therefore, the optimization of bathymetric data acquisition in terms of accuracy and cost-effectiveness is economically important.

One of the most efficient and accurate bathymetric survey methods today is the one using the multibeam echosounder. The device enables obtaining datasets of

J.-S. Pan, S.-M. Chen, N.T. Nguyen (Eds.): ACIIDS 2012, Part III, LNAI 7198, pp. 456–465, 2012.

measurement points covering a strip of seabed called the *profile*, corresponding to the straight route covered by the survey vessel [1]. The points are situated on lines, referred to as *measurement lines*. The survey of an entire area usually necessitates the registration of many profiles, designed as a series of sections, whose arrangement depends on the surface shape. Eventually, the distribution of measurement points depends on the seabed relief, or shape, and on echosounder parameters (e.g. the number and angle of beams or impulse frequency) and on bathymetric survey parameters, such as vessel speed or adopted arrangement of profiles.

As a rule, the recording of area seabed data by the multibeam echosounder brings an enormous number of points having irregular distribution in space. Such data, due to their size and distribution, cannot be directly used for, say, visualization or analysis. Bathymetric data are transformed into more organized data structures, such as TIN (triangulated irregular network) and GRID (regular square net), called digital terrain models (DTM). Interpolation methods of GRID modeling based on irregular measurement data can be found in various publications, e.g. [2-7]. However, these methods have not been examined taking into account the specific data from the multibeam echosounder, particularly with the aim of estimating the influence of echosounder or survey parameters on model accuracy.

This article presents research done to define the effect of marine survey parameters determining the density of measurement points (vessel speed, beam angle, number of echosounder beams, distances between profiles) on the accuracy of created seabed models.

1.1 The Importance of Accuracy in Seabed Modeling

The most important criterion in seabed modeling is accuracy, expressed as errors: differences between each point of the DTM and the corresponding point on the real bottom. Bathymetric survey should be performed in a manner allowing to estimate each error value, thus enabling the overall estimation of the created model accuracy. The need to estimate overall DTM error results from the requirement of high reliability that maps must have, and maximum error values are specified by the International Hydrographic Organization (IHO) regulations [8].

One essential problem in the process of creating a DTM is the impossibility of precise estimating of model accuracy. This is due to the fact that the actual shape of the surveyed seabed surface is not known, consequently we cannot compare the created model with the original surface. Practically, the determination of accuracy consists in the estimation of, then summing up the errors that occur in each stage of modeling. It is commonly adopted that the depth error corresponds to the measuring accuracy of the instrument, specified by the instrument maker. The other error components are often neglected on the assumption that they are so small that they hardly affect the total error.

The authors' extensive research has resulted in the development of the methodology for DTM accuracy testing called *virtual survey*, using the authors' multibeam echosounder simulator ([9-10]). The simulator takes into account the most essential parameters of the survey (reference seabed model, vessel speed, arrangement

of profiles, beam angle, number of beams, echosounder working frequency), which enables examining the errors of the obtained DTMs depending on the selection of these parameters.

2 The Influence of Marine Survey Parameters on Measurement Points Density

Source data consist of millions of points collected during one measurement session by hydroacoustic instruments. When the multibeam echosounder is in operation, many single points lying on a measurement line perpendicular to the vessel track are recorded periodically. The number of points depends on the *number of beams* transmitted by the device, and the length of the measurement line depends on the *beam width* setting. The intervals between subsequent measurement lines depend on echosounder *frequency* (which is usually constant for the given device, e.g. 10Hz) and on vessel *speed*. These factors essentially affect the number and distribution of measurement points. The methodology of survey by multibeam echosounder is described in several works, e.g. [11]. In most cases, a series of parallel longitudinal profiles is performed, only in some justified cases transverse or diagonal profiles as made as well. To ensure full survey coverage of the seabed, the *overlapping* of profiles is usually applied (leading to local increase in the number of measurement points).

Planning marine survey one has to take into consideration high accuracy requirement of modeling on the one hand, and costs of survey works on the other hand. The reduced speed of the survey vessel, additional transverse profiles, increased profile density or smaller beam angle will improve model accuracy, but they will significantly extend the survey duration. It is therefore purposeful to assess what effect these parameters have on the accuracy and survey costs. To make such assessment possible, the authors use reference GRID models of high resolution (0.1m × 0.1m) and the multibeam echosounder simulator that performs virtual measurements. From data thus obtained for various combinations of input parameters, test surface models are created and directly compared with the reference surface. Consequently, it is possible to determine the errors produced in data acquisition and GRID structure modeling.

2.1 Reference Surfaces

The tests included two reference surface areas acquired from real measurement data, collected by Szczecin Maritime Office vessels in the areas of the Piastowski Canal and Pomorska Bay. These waters have diversified bottom shape (Fig. 1):

- *anchorage* – flat bottom, without sudden depth changes,
- *swinging area* – varied bottom surface, with sudden depth changes and untypical forms due to dredging works.

Each surface is recorded in the GRID structure with a 0.1m × 0.1m and 1m × 1m resolutions and covers a square area 200m × 200m. The 0.1m × 0.1m model can be said to have high resolution, while 1m × 1m resolution is typical in the practice of marine surveys.

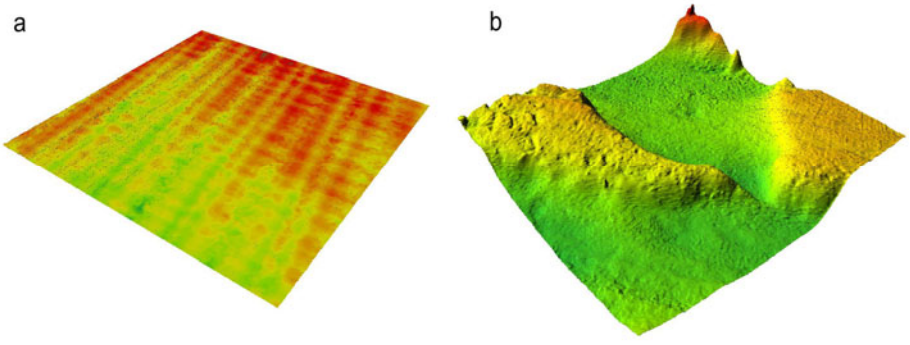


Fig. 1. Reference surfaces: anchorage (a), swinging area (b)

2.2 Virtual Survey

The simulator is used to simulate the survey vessel movement along a preset route and collection of measurement points (x,y,z) from the indicated reference surface. Simulations are varied in terms of vessel parameters and echosounder operating characteristics. The simulator features a measurement error generator, developed from data obtained by examining the real error distribution of an echosounder Simrad EM3000 [11].

2.3 Testing Procedure

Five virtual surveys varying in vessel and/or echosounder parameters were performed for each reference surface. The resulting data represented measurement points of diversified density. The surveys characterized by a specific set of input parameters, are named with terms corresponding to the density of measured points: *very low*, *low*, *regular*, *high*, *very high*. The variable parameters of surveys, the number and density of acquired measurement points are shown in Table 1, while the measurement points distributions for each combination of parameters are illustrated in Fig. 2.

Bathymetric data from each of the virtual surveys were used to build GRID models with the two resolutions: 1m × 1m and 0.1m × 0.1m, using the kriging method of interpolation. The ranges and resolutions of the grids were the same as in the reference models, so it was possible to directly compare them (in the 1m × 1m grids their subsequent nodes corresponded to every tenth node in the reference model).

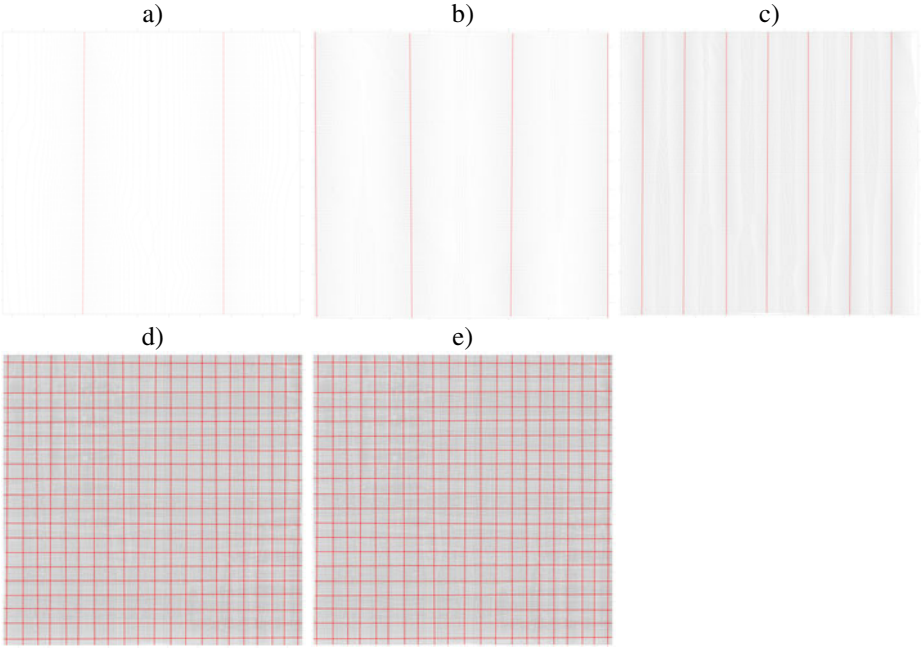


Fig. 2. Distributions of measurement points acquired in each survey (part of the *swinging area* surface): a) very-low, b) low, c) regular, d) high, e) very-high

Table 1. Survey parameters and the obtained density of measurement points

Survey	Speed [kn]	Angle [deg]	Profiles	Over-lapping [%]	Number of beams	Freque-ncy [Hz]	Number of points *anchorage*	Average density (pt/m²)	Number of points *swinging area*	Average density (pt/m²)
Very low	10	150		No	63	10	73100	1.8	51202	1.3
Low	7.5	130		No	127	10	315660	7.9	270982	6.8
Regular	5	110		15	127	10	804486	20.1	709609	17.7
High	3.5	110		33	127	10	2992404	74.8	2474541	61.8
Very high	2	90		50	127	10	10009102	250.2	10167391	254.2

The analysis of the results aimed at determining the maximum error, mean error, standard deviation and the error value at 95% and 99% confidence levels. The analysis of these quantities enables assessing the accuracy of the created model.

3 The Results

The results obtained for two reference surfaces and for two resolutions of of the GRID structure are presented in Figures 3-6.

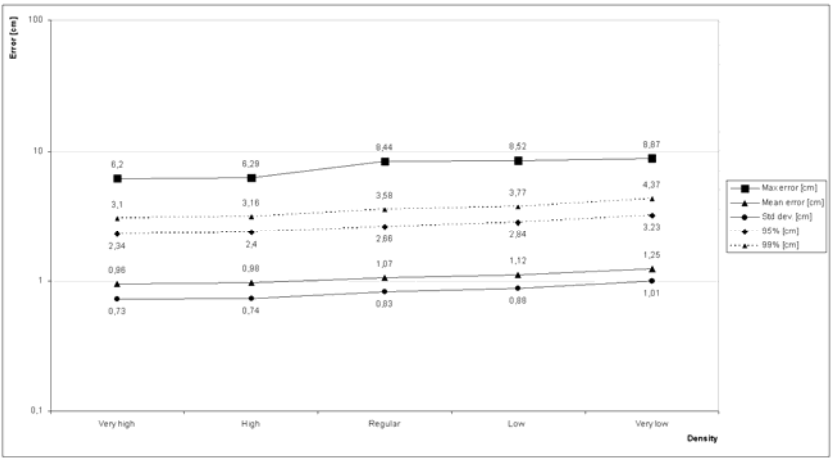


Fig. 3. The results for the surface of *anchorage* (0.1m × 0.1m)

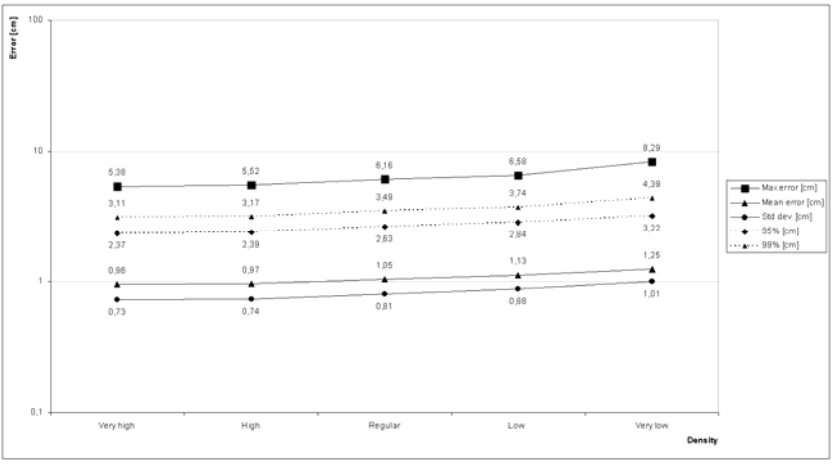


Fig. 4. The results for the surface of *anchorage* (1m × 1m)

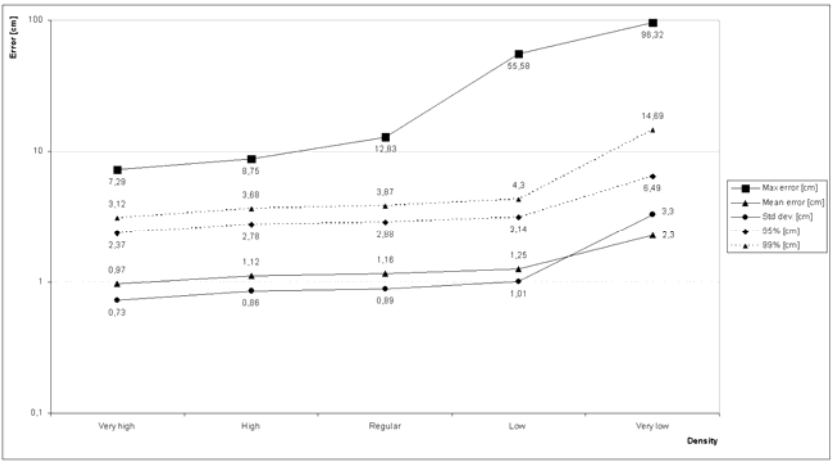


Fig. 5. The results for the surface of *swinging area* (0.1m × 0.1m)

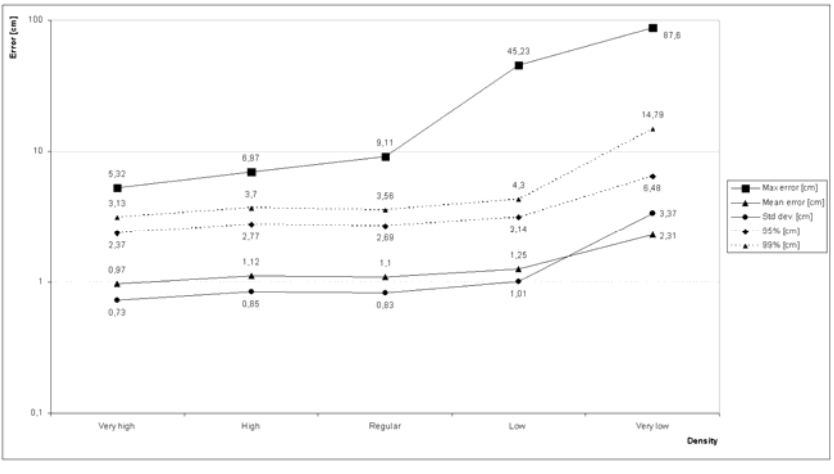


Fig. 6. The results for the surface of *swinging area* (1m × 1m)

The following observations can be formulated from the analysis of the survey results:

- Increase in data density leads to higher DTM accuracy,
- Four out of five examined parameter sets (except for *very low*) enabled collecting a sufficient amount of data in order to build a high quality model. It was assumed that for errors of less than 5cm at the 99% confidence level the DTM is of high quality. For the four surveys (*low*, *regular*, *high*, and *very high*) the mean error was less than 2cm, and 99% of the points were modelled at 3-5 cm accuracy.
- Increase in data density leads to higher DTM accuracy.

- All modeled surfaces satisfied IHO requirements (error not larger than 20cm for depths characterizing the models.
- Significant increase of the measurement points density (e.g. tenfold increase for *regular* and *very high* surveys) only slightly improves the modelling accuracy (error decreased by approx. 0.5cm). Therefore, it does not seem economical to perform very accurate measurements (by slowing down the the vessel, concentrating the profiles, additional transverse profiles). In the above case the vessel had to cover a distance five times longer, and the measurement time increased 18 times. In spite of such considerable increase in the data density, the model accuracy rose only by 0.05%. Thus the question is: will a substantial reduction of the number of measurement points cause a substantial reduction of the created model quality?
- For the surveys *regular* and *very low* (15-fold reduction of measurement points density, 2.5 times shorter route, nearly 4 times shorter time of survey), the results are visibly diversified, depending on the bottom shape:
  1. For the flat surface (anchorage) the mean error of the model increases by 0.6 cm, and the error at 99% confidence level is about 5cm. The obtained model can be regarded as correct.
  2. For a varied surface (swinging area) the mean error of the model increases by 10cm, and the error at 99% confidence level is 15cm. Such surface, still satisfying IHO requirements, seems to be rather inaccurately modeled, as many small forms get smoothened and deformed. The authors are of the opinion that the corresponding combination of survey parameters results in insufficient data collection for a model to have satisfactory accuracy.
- The analysis shows that the survey parameters have varying effect on the density and spatial distribution of measurement points:
  1. Relationship between vessel speed and point density is linear, and actually this parameter slightly affects model accuracy.
  2. Profile density or additional transverse profiles substantially increase data density (the density doubles in areas where data from profiles overlap each other), while their effect on the model accuracy is moderate.
  3. The beam angle does not affect the amount of collected data, but it determines points distribution and density. At the ends of the echosounder beam scope the point density is lower than in the middle (distances between adjacent points can exceed even 1m for wide beams). In these areas considerable errors occur due to the averaging of data in the process of modeling. This parameter has the greatest effect on modelling accuracy. It seems purposeful to set the beam angle in the 110-130 degree range.

3.1 Conclusions

The following conclusions can be formulated from the gathered data:

1. Significant increase in the survey vessel speed reduces the amount of measurement points collected, but their number is sufficient for the creation of accurate DTM models, (provided that the echosounder frequency is not less than 10Hz, and the DTM resolution not higher than 0.1m × 0.1m). To optimize survey costs, the vessel may survey an area at a relatively high speed, e.g. 10 knots.
2. Where the seabed surface is considerably varied, and highest quality models are required, it is recommended to make measurements at lower speed (2 - 5 knots), which will add to the model accuracy by 50%, even 100% (as compared to the survey vessel speed of 10 to 16 knots). Besides, by using additional transverse profiles we can raise the model accuracy by approx. 20%.
3. In the surveys done the greatest errors occurred at the edges of areas measured, therefore it is recommended to take measurements of an area slightly larger than the one to be modeled (sufficient margin is a distance equal to a few nodes of the ultimate GRID network).
4. Models built on the basis of two different GRID networks: 1m × 1m and 0.1m × 0.1m, yielded nearly the same accuracy of depth modeling. We can then state that based on the same source data, models with various GRID resolution have similar accuracy of depth mapping. This property, however, requires further research.

4 Summary

The accuracy tests of seabed surface models created from datasets produced by the multibeam echosounder show that even the modest coverage of an area with measurement points, namely 10 points per square meter is sufficient for building a highly accurate model (error in cm is a one digit value). Bearing this in mind, to run economically optimized seabed surveys, the vessel can move at higher speeds, and the arrangement of profiles should meet this guideline: 100% bottom coverage with least possible number of profiles. Only in areas where highest possible accuracy is required (surveys preceding marine construction projects) it is recommended to survey the seabed at slow vessel speed (e.g. 3 knots) and by performing additional transverse profiles. Then the accuracy of the DTM can be even twice higher.

All the models built within the research had accuracy satisfying IHO standards. Considering the estimated mean error of measurement it can be stated that for little varied flat surfaces it was about 1cm, while for more varied surfaces - 1.5cm. It should be noted that this level of accuracy is sufficient for most DTM applications, including bathymetric charts. Only measurements bringing the lowest density of measurement points (survey: *very low*) can be regarded as insufficient if one intends to produce a model with accuracy meeting IHO standards.

The research described herein took into consideration only the echosounder measurement error and the GRID modeling error relating to the distribution and density of measurement points. In case of results of an actual survey the DTM error can be also affected by such factors as: accuracy of positioning systems, stability

systems and errors due to the echosounder design and mounting. The authors intend to verify the research data presented in this article by performing real survey of the chosen seabed surface, applying various survey parameters.

Acknowledgements. This work was supported by the Polish Ministry of Science and Higher Education through the grant no. N N526 073038.

References

1. Maleika, W., Palczynski, M., Frejlichowski, D., Stateczny, A.: Analysis of survey data collected using Simrad EM3000 multibeam echosounder. Metody Informatyki Stosowanej, PAN o. Gdansk 4(25), 55–64 (2010) (in Polish)
2. Calder, B.R., Mayer, L.A.: Automatic processing of high-rate, high-density multibeam echosounder data. Geochemistry Geophysics Geosystems 4(6) (2003)
3. Dinn, D.F., Loncarevic, B.D., Costello, G.: The effect of sound velocity errors on multi-beam sonar depth accuracy. In: Proceedings of MTS/IEEE Conference on Challenges of Our Changing Global Environment, Oceans 1995, vol. 2, pp. 1001–1010 (1995)
4. Hamilton, E.L.: Geoacoustic modeling of the sea floor. Journal of the Acoustical Society of America 68(5), 1313–1340 (1980)
5. Lubczonek, J., Borkowski, M.: Comparative analysis of digital seabed models prepared from single and multibeam sounding data. Polish Journal of Environmental Studies 19(5), 1039–1043 (2010)
6. Gao, J.: Resolution and accuracy of terrain representation by grid. International Journal of Geographical Information Science 11(2) (2001)
7. Stateczny, A.: (red.): The methods of the comparative navigation. In: Gdanskie Towarzystwo Naukowe, Gdansk (2004) (in Polish)
8. International Hydrographic Organization: IHO standards for hydrographic surveys, Publication No. 44, 4th edn. (1998)
9. Maleika, W., Palczynski, M.: Development of a virtual multibeam echosounder. Biuletyn WAT XL (3(663)), 215–225 (2011) (in Polish)
10. Maleika, W., Palczynski, M.: Virtual multibeam echosounder in investigations on sea bottom modeling. Metody Informatyki Stosowanej, PAN o. Gdansk 4, 111–120 (2008)
11. Maleika, W., Pałczyński, M., Frejlichowski, D.: Multibeam Echosounder Simulator Applying Noise Generator for the Purpose of Sea Bottom Visualisation. In: Maino, G., Foresti, G.L. (eds.) ICIAP 2011, Part II. LNCS, vol. 6979, pp. 285–293. Springer, Heidelberg (2011)
12. 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)
13. 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)

Interpolation Methods and the Accuracy of Bathymetric Seabed Models Based on Multibeam Echosounder Data

Wojciech Maleika, Michal Palczynski, and Dariusz Frejlichowski

West Pomeranian University of Technology, Szczecin
Faculty of Computer Science and Information Systems,
Zolnierska 52, 71-719, Szczecin, Poland
{wmaleika,mpalczynski,dfrejlichowski}@wi.zut.edu.pl

Abstract. In order to make reliable sea bottom visualizations and analyses, accurate bathymetric models are necessary. The authors analyzed the process of GRID model creation using multibeam echosounder data and pointed interpolation methods as important sources of models' errors. In order to assess the accuracy of the model, the simulation technique named *virtual survey* was applied. A wide range of interpolation methods was examined. The results show significant differences between these methods in terms of accuracy and effectiveness. The choice of the best interpolation methods for various cases is suggested.

Keywords: Surface modeling, interpolation methods, multibeam echosounder.

1 Introduction

The exploitation of water areas, maritime transportation as well as the exploration of the seabed or sub-bottom resources, requires detailed knowledge of the depth. This information is usually processed and presented using computer imagery methods, which tend to give increasingly wider possibilities of data analysis. Unlike the land areas, where thanks to global positioning systems or other tools of geodesy it is possible to measure the accurate elevation of every point, the measurement of water depth is much less accurate and much more expensive. In addition, it has to be updated because of rapid changes of the seabed shape. The optimization of the process of gathering bathymetric data may be of major economical importance.

Nowadays, one of the most effective and accurate methods of depth measurement is the bathymetric survey using multibeam echosounder. A large number of measurements is collected, where the measurement points cover the strip of the bottom along the ship's route, usually a straight line named *profile*. Complete bathymetry of the sea area requires making many profiles. Their number and composition depends on the shape of the area. Measurement points always form lines, orthogonal to the route [1]. An example of bathymetric data is shown in Fig. 1. The distribution of measurement points depends on the bottom shape, echosounder parameters, such as the number and angle of the beams and survey parameters, such as ship's speed or the composition of profiles. Bathymetric survey using multibeam

J.-S. Pan, S.-M. Chen, N.T. Nguyen (Eds.): ACIIDS 2012, Part III, LNAI 7198, pp. 466–475, 2012.

echosounder usually results in collection of huge number of points distributed irregularly in space. Such data cause difficulties in analysis, processing and visualization because of the size and distribution. Usually, survey data are transformed using interpolation methods to much more organized models: TIN (triangulated irregular network) or GRID (regular square network), named digital terrain models (DTM). These models are a good basis for analyses, visualizations and charting. The methods for building DTMs using multibeam echosounder data are described in [2-7]. In this article, an analysis of factors decreasing the accuracy of models is described. The results of investigation of the interpolation methods influence on the GRID accuracy using multibeam echosounder data are shown.

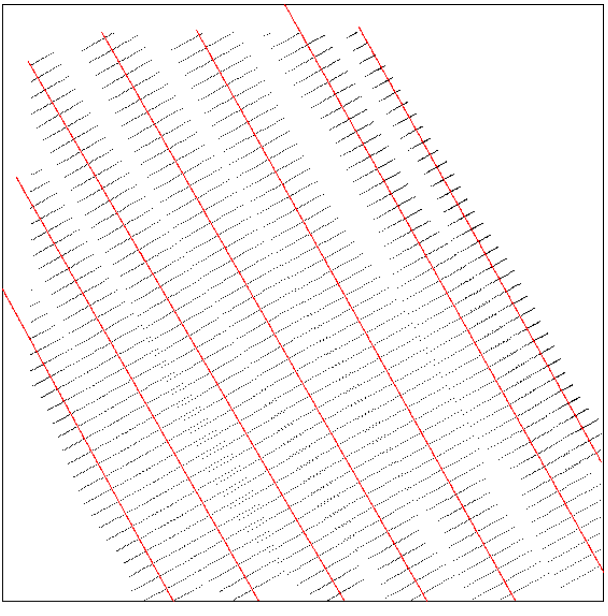


Fig. 1. Orthogonal projection of bathymetric data. Red lines show the profiles of the survey.

1.1 Accuracy in Seabed Modeling

The most important criterion in seabed modeling is accuracy, expressed as errors: differences between each point of the DTM and the corresponding point on the real bottom. Bathymetric survey should allow the assessment of the model accuracy. The general error of the modeling is the result of errors made during the subsequent stages of DTM building:

- errors of the measurement device (depend on depth and type of seabed, model of the device – usually specified by the producer)
- errors caused by survey parameters (speed, profiles composition, echosounder parameters – difficult to assess, usually ignored),
- position errors (depend on positioning system),

- errors caused by the interpolation process (so far ignored as they are difficult to assess),
- errors caused by smoothing (ignored so far).

The necessity of error assessment for DTM is caused by the requirements of high reliability of maps. Maximum error values are given by the IHO - International Hydrographic Organization ([8]). However, the main problem in the bathymetric work is the inability to accurately calculate the error value for the resulting DTM, because the real shape of the bottom is unknown, so there is no reference model. The total error should be determined by assessing and adding together errors of subsequent stages of the DTM modeling. Usually, the nominal error of the measurement device is taken as the total error, since other possible errors are assumed to be irrelevant, which is oversimplification. The authors examine errors of GRID models caused by the interpolation process and compare various methods in terms of accuracy and effectiveness.

2 Research on the Accuracy of Interpolation Methods

In the process of DTM modeling (creating GRID structure), new points are created according to one of interpolation methods. The results of various methods (and the DTM errors) are different. The choice of the most accurate method is often difficult because it depends on the bottom shape, density and distribution of measurement points.

Many research works have been done in the area of interpolation methods accuracy in connection with sea bottom modeling. Many algorithms and their parameters were examined, taking into account input data characteristics (method of collecting, number of measurements, type of the area, seabed shape) and DTM model (GRID, TIN) ([2,3,9]). The conclusions show that if the data consist of a great number of measurement points (at least 1000), the algorithms Minimum Curvature and Triangulation with Linear Interpolation give the best results: small errors and short time of data processing. The Kriging and Radial Basis Function methods yield similar accuracy but much lower efficiency ([7]). There are few works describing the investigations using huge data sets (millions of measurement points), which are typical for multibeam echosounder bathymetry. In most of all those works, interpolation methods were compared on the basis of synthetic test surfaces, built using mathematical formulae and the random measurement points were chosen. Although such approach allows to calculate exact error values, there are some disadvantages:

- shapes of test surfaces (usually defined as a combination trigonometric functions) don't meet the real bottom characteristics,
- random distribution of test measurement points differs from the distribution typical for the data collected by echosounder,
- number of points is not under control and often too little, while it's very important factor for the choice of the interpolation method,

– resolutions of the models (number of GRID nodes per meter) is usually low (1 node per 1-2m), thus the accuracy of models including small features on seabed can't be examined.

In order to eliminate these problems, a different idea was developed [10]. First, various reference surfaces were prepared in the form of high resolution GRID models, based on real data. Then, simulated (virtual) surveys were done resulting in data sets similar to the real bathymetry. These data were the basis for the test DTM development. This procedure allows to calculate DTM errors directly as the difference between the test and reference models.

2.1 Reference Surfaces

The research was done using four surfaces prepared using real bathymetric data collected by Szczecin Maritime Office.

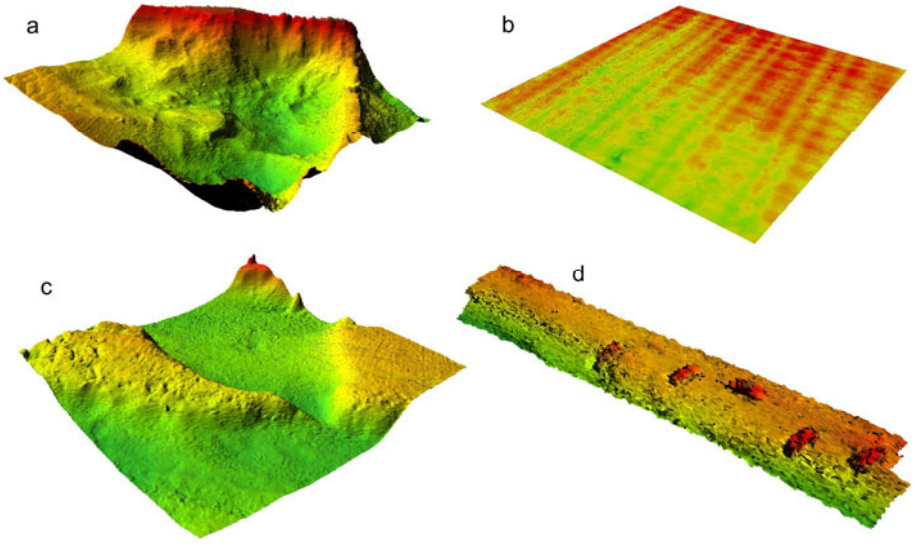


Fig. 2. Test surfaces: gate (a), anchorage (b), swinging area (c), wrecks (d)

The surfaces differ significantly in their shapes:

– *gate* – significant but gradual depth changes,
– *anchorage* – almost flat, only very little shapes are present,
– *swinging area* – very varied surface, sudden holes and slopes,
– *wrecks* – almost flat surface with a few atypical objects: car wrecks.

A GRID model was built for each of the surfaces. Since the size of each area was 200m × 200m and the resolution was 0.1m × 0.1m, the surfaces can be assumed as high resolution reference models. The reference surfaces are shown in Fig. 2.

2.2 The Research Procedure

In order to collect bathymetric data, needed for the examination of interpolation methods, virtual surveys were done over reference surfaces using the multibeam echosounder simulator developed by the authors ([11,12,13]). It simulates the motions of the survey vessel and the operation of the echosounder, which results in calculating virtual measurement points (x,y,z) on the input surface according to the ship's route and echosounder parameters. The parameters of virtual survey:

- speed: 5 knots,
- number of beams: 127,
- angle of beams: 110º,
- impulse frequency: 10Hz,
- profiles composition: parallel, overlapping: 20%,
- surface coverage: 100%.

The data generated by virtual surveys made up a basis for the examination of interpolation methods. A large group of these methods, recommended in the literature, was chosen, some of them versions of the same method. The differences between the versions consisted mostly in the search radius, where the specific values were chosen according to the GRID resolution. Additional smoothing options were switched on, if possible. The list of investigated interpolation methods (11 methods and their 31 variants):

- Inverse Distance to a Power (8 variants),
- Kriging (6 variants),
- Minimum Curvature,
- Modified Shepard's Method (5 variants),
- Natural Neighbor,
- Nearest Neighbor (3 variants),
- Simple Planar Surface,
- Triangulation with Linear Interpolation,
- Moving Average (3 variants).
- Local Polynomial (3rd degree),
- Radial Basis Function.

For each method and each of the reference surfaces, the test GRID model was calculated and the time of operation was measured. The size and resolution of each model was the same as those of the reference surface. Then, the difference between corresponding test and reference DTMs was calculated. The resulting DTM formed the *error surface,* representing the distribution of error values. Next, for that error surface the maximum error, mean error and standard deviation were calculated.

Two versions of the results analyses were applied:

v1: Accuracy and time analysis – the conclusions affect processing of large surface areas, where the time can be significant.

v2: Accuracy analysis – the time of the processing is ignored.

The calculated statistical values were transformed to standardized scores (0-100 pts). The total score for each interpolation method was calculated as the sum

of the individual scores multiplied by the weights expressing the importance of the corresponding parameter. The scores and weights for each parameter and analysis version are specified in Table 1.

Table 1. Scores and weights for the results of the DTM research

Parameter	min. score: 0 pts	max. score: 100 pts	weight (v1)	weight (v2)
processing time	≥ 1000 s	0 s	0.2	0.0
maximum error	≥ 100 cm	0 cm	0.1	0.1
mean error	≥ 20 cm	0 cm	0.4	0.6
standard deviation	≥ 20 cm	0 cm	0.3	0.3

The hardware environment used for the research:

- Computer: Sony Vaio VPCZ11X9E (Intel Core I5 2.4 GHz, 4GB RAM, HDD Samsung SSD),
- Operating system: Windows 7 64-bit,
- Gridding software: Golden Software Surfer 9.0.

3 The Results

All chosen interpolation methods were examined using each of four reference surfaces, according to the procedure described in section 2.2. The final scores for each method were calculated by averaging scores obtained for each reference surface. The comparison of the results is shown in Table 2 (analysis v1 – including time consideration) and Table 3 (analysis v2 – accuracy only). The abbreviations used are listed below:

TLI – Triangulation with Linear Interpolation,
IDP – Inverse Distance to a Power,
MA – Moving Average,
MSM – Modified Shepard's Method,
NN – Nearest Neighbor,
SPS – Simple Planar Surface,
ME – mean error,
ne – nuggets effect.
K – Kriging,
LP3 – Local Polynomial (3rd degree),
MC – Minimum Curvature,
NatN – Natural Neighbor,
RBF – Radial Basis Function,
MaxE – maximum error,
SD – standard deviation,

Table 2. Final scores of each interpolation method (analysis v1: accuracy and time)

No	Method	Smoothing method	Search radius	Achorage	Swinging area	Gate	Wrecks	Average
1	MA	No	1x1	96.97	91.72	84.81	70.12	**85.9**
2	IDP	No	1x1	95.09	90.57	82.48	73.74	**85.47**

Table 2. (*continued*)

3	IDP	No	1x1	90.7	89.34	81.99	79.5	**85.38**
4	IDP	factor=1	1x1	95.1	90.18	81.99	73	**85.07**
5	IDP	factor=5	1x1	95.09	90.16	81.97	72.97	**85.05**
6	K	factor=10	1x1	86.69	85.38	78.75	75.44	**81.56**
7	TLI	factor=5	-	89.54	83.83	63.48	84.29	**80.29**
8	NN	factor=5	1x1	85.96	82.94	75.59	76.58	**80.27**
9	IDP	No	5x5	90.02	88.58	63.25	76.69	**79.64**
10	IDP	No	5x5	94.47	88.67	62.65	69.4	**78.8**
11	IDP	No	10x10	90.08	88.57	58.94	72.44	**77.51**
12	IDP	ne = 1.1	10x10	94.44	88.67	58.17	65.2	**76.62**
13	NN	ne = 3.3	5x5	85.96	82.95	61.22	74.91	**76.26**
14	K	ne = 10.10	5x5	86.59	82.32	64.96	68.96	**75.71**
15	K	No	1x1	72.22	75.74	68.69	74.8	**72.86**
16	NN	No	10x10	85.96	82.95	56.19	65.35	**72.61**
17	K	No	10x10	87.28	82.63	53.52	63.02	**71.61**
18	MC	No	-	87.48	78.4	24.57	82.72	**68.3**
19	K	factor=0.1	5x5	70.21	69.44	48.52	70.7	**64.72**
20	K	factor=2	10x10	70.21	69.55	48.22	66.43	**63.6**
21	MA	No	5x5	95.61	73.22	31.56	19.02	**54.85**
22	RBF	No	1x1	35.22	59.21	63.55	59.92	**54.47**
23	MSM	No	5x5	48.99	70.76	34.67	19.92	**43.58**
24	MSM	No	30x30	48.99	70.76	34.66	19.92	**43.58**
25	MSM	No	1x1	49.01	70.53	34.67	19.94	**43.54**
26	MSM	No	5x5	48.82	71.92	32.44	19.58	**43.19**
27	NatN	No	-	88.51	0	0	83.64	**43.04**
28	SPS	No	-	92.59	19.96	19.97	19.98	**38.13**
29	MSM	No	5x5	34.46	60.16	19.6	11.22	**31.36**
30	LP3	No	1x1	60.85	15.2	16.49	16.35	**27.22**
31	MA	No	30x30	74.14	0	3	16.52	**23.42**

Table 3. Final scores of each interpolation method (analysis v2: accuracy only)

No	Method	Smoothing method	Search radius	Achorage	Swinging area	Gate	Wrecks	Average
1	K	No	1x1	95	92.34	84.27	78.44	**87.51**
2	IDP	No	1x1	96.23	91.05	82.52	72	**85.45**
3	IDP	No	1x1	90.93	89.22	81.71	78.75	**85.15**
4	IDP	factor=1	1x1	96.25	90.61	81.96	71.1	**84.98**
5	IDP	factor=5	1x1	96.25	90.57	81.91	71.06	**84.94**
6	K	factor=10	1x1	87.66	86.63	80.13	82.41	**84.21**

Table 3. (*continued*)

7	MA	factor=5	1x1	96.29	90.55	83.15	65.84	**83.96**
8	K	factor=5	5x5	95.99	91.94	72.87	74.99	**83.94**
9	K	No	10x10	96.31	90.95	63.69	72.56	**80.88**
10	IDP	No	5x5	91.03	89.36	63.71	78.74	**80.71**
11	K	No	5x5	87.65	86.63	64.92	82.4	**80.4**
12	K	ne = 1.1	10x10	87.65	86.63	64.52	82.4	**80.3**
13	IDP	ne = 3.3	10x10	91.06	89.36	61.8	78.74	**80.24**
14	IDP	ne = 10.10	5x5	96.35	89.9	63.15	70.59	**80**
15	IDP	No	10x10	96.35	89.9	60.95	70.59	**79.45**
16	TLI	No	-	86.95	81.26	59.93	81.9	**77.51**
17	NN	No	1x1	82.49	79.47	71.58	73.56	**76.77**
18	NN	No	5x5	82.49	79.47	56.87	73.56	**73.09**
19	NN	factor=0.1	10x10	82.49	79.47	53.03	73.56	**72.13**
20	RBF	factor=2	1x1	51.38	76.14	76.74	76.19	**70.11**
21	MC	No	-	85.63	76.83	8.97	81.24	**63.17**
22	MA	No	5x5	95.54	69.95	18.27	0	**45.94**
23	NatN	No	-	88.02	0	0	84.11	**43.03**
24	MSM	No	5x5	44.25	67.62	22.56	0	**33.61**
25	MSM	No	30x30	44.25	67.62	22.56	0	**33.61**
26	MSM	No	1x1	44.28	67.38	22.56	0	**33.56**
27	MSM	No	5x5	44.16	69.05	19.35	0	**33.14**
28	MSM	No	5x5	43.77	69.21	19.35	0	**33.08**
29	MA	No	30x30	92.54	0	0	0	**23.13**
30	SPS	No	-	90.61	0	0	0	**22.65**
31	LP3	No	1x1	63.36	0	0	0	**15.84**

The analysis of the results allows to choose the group of interpolation methods which process huge data sets significantly better than the others. These are: *kriging*, *inverse distance to the power* and *moving average*. *Triangulation with linear interpolation* turned out to be slightly worse in terms of accuracy but very fast. The other methods generate much greater error values, so they are regarded as unsuitable for DTM creation using large data sets.

For the methods where the search radius was tested, better results were observed for smaller radius – the accuracy slightly increased and the time of processing was significantly reduced. It can be explained by the number and density of measurement points generated by the multibeam echosounder. We can usually find tens of points per $1m^2$, which is definitely enough for high density DTM development. The results show clearly the advantage of the variants with the search radius of 1m over those where longer radius values were chosen (5, 10, 30).

Additional smoothing resulted in a slight increase of the accuracy, but only for low values of the smoothing factor. Too high values resulted in the absence of small objects in the model.

If the accuracy and effectiveness are considered (analysis v1), the best method was the m*oving average* with the search radius set to 1m. Very good results were also

obtained for *inverse distance to a power* and *kriging*, with search radius set to 1m, and smooth factor (*inverse distance to a power*) equal to 1, or nuggets effect (*kriging*) equal to 1.1. Although *kriging* ensures slightly higher accuracy, this method ranks lower due to the processing time, up to 10 times longer than in other methods. However, for isolated surfaces containing small objects or very varied forms (*swinging area* or *wrecks* models), *inverse distance to a power* method could give better results.

If the accuracy of the model is the only criterion (analysis v2), the best method is *kriging* with the radius set to 1m and nuggets effect parameter equal to 1.1. If the surface contains small objects (for example *wrecks* model), it is better to avoid additional smoothing. Good results were achieved also for variants of *inverse distance to a power. Moving average* ranked seventh. It should be mentioned that the differences between the results of those methods are very small: mean error for *kriging* (1st place) was 1.41cm (processing time: 370s), while for *moving average* (7th) the mean error was 1.99cm (processing time: 2s). Thus the difference between the errors is small but the time differs significantly.

Triangulation with linear interpolation is the fastest method giving relatively good accuracy. This method, as well as *moving average* could be useful for quick, provisional DTM development.

One can note high values of the maximum error, up to 1m. It should be mentioned that such big errors are very rare (a few values out of 400000 nodes) and their positions are usually on the edges of models. Typical example of error distribution is presented in fig. 3.

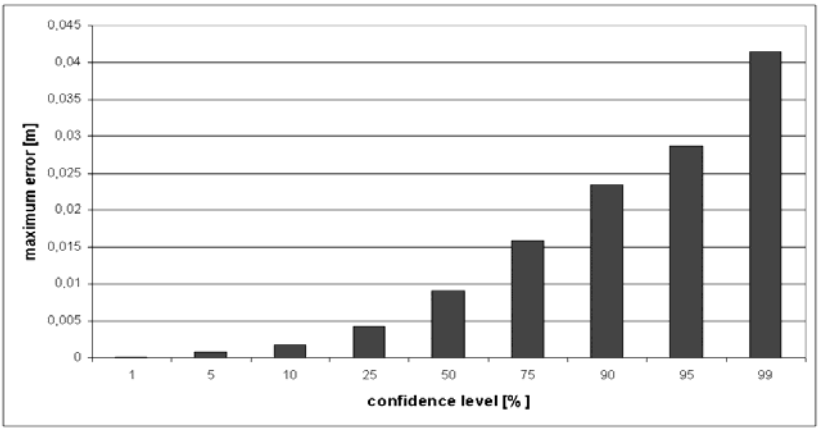


Fig. 3. Error distribution obtained for kriging (search radius: 1m, nuggets effect: 1.1)

4 Summary

The examination of the accuracy of DTM created using multibeam echosounder data has shown differences in error values and processing time for interpolation methods

under consideration. The methods such as *kriging*, *inverse distance to a power* and m*oving average* yielded significantly better results than the other methods. If the processing time is important or input data set is large (millions of points), *moving average* is recommended. If the accuracy is the only criterion, *kriging* should be used.

Another conclusion from the results analysis concerns the level of mean error caused by the interpolation process. For smooth surfaces its value does not exceed 5 cm, while for more varied surfaces it can reach 10cm (at confidence level of 99%). In comparison with the IHO requirements, according to which the error should not exceed 20-30cm for surfaces used for the examination, the obtained values meet the relevant standards. However, they are significant and cannot be ignored.

Acknowledgements. This work was supported by the Polish Ministry of Science and Higher Education through the grant no. N N526 073038.

References

1. Maleika, W., Palczynski, M., Frejlichowski, D., Stateczny, A.: Analysis of survey data collected using Simrad EM3000 multibeam echosounder. Metody Informatyki Stosowanej, PAN o. Gdansk 4(25), 55–64 (2010) (in Polish)
2. Calder, B.R., Mayer, L.A.: Automatic processing of high-rate, high-density multibeam echosounder data. Geochemistry Geophysics Geosystems 4(6) (2003)
3. Dinn, D.F., Loncarevic, B.D., Costello, G.: The effect of sound velocity errors on multi-beam sonar depth accuracy. In: Proceedings of MTS/IEEE Conference on Challenges of Our Changing Global Environment, Oceans 1995, vol. 2, pp. 1001–1010 (1995)
4. Hamilton, E.L.: Geoacoustic modeling of the sea floor. Journal of the Acoustical Society of America 68(5), 1313–1340 (1980)
5. Gao, J.: Resolution and accuracy of terrain representation by grid DEMs at a micro-scale. International Journal of Geographical Information Science 11(2), 199–212 (2001)
6. Stateczny, A. (red.): The methods of the comparative navigation. In: Gdanskie Towarzystwo Naukowe, Gdansk (2004) (in Polish)
7. International Hydrographic Organization: IHO standards for hydrographic surveys, Publication No. 44, 4th edn. (1998)
8. Hammerstad, E., Asheim, S., Nilsen, K., Bodholt, H.: Advances in multibeam echosounder technology. In: Proceedings of Engineering in Harmony with Ocean, Oceans 1993, vol. 1, pp. I482–I487 (1993)
9. Maleika, W., Palczynski, M.: Virtual multibeam echosounder in investigations on sea Bottom modeling. Metody Informatyki Stosowanej, PAN o. Gdansk 4, 111–120 (2008)
10. Maleika, W., Palczynski, M.: Development of a simulator of multibeam echosounder. Biuletyn WAT (3(663)), 215–225 (2011) (in Polish)
11. Maleika, W., Pałczyński, M., Frejlichowski, D.: Multibeam Echosounder Simulator Applying Noise Generator for the Purpose of Sea Bottom Visualisation. In: Maino, G., Foresti, G.L. (eds.) ICIAP 2011, Part II. LNCS, vol. 6979, pp. 285–293. Springer, Heidelberg (2011)
12. 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)
13. 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)

Quantitative Measurement for Pathological Change of Pulley Tissue from Microscopic Images via Color-Based Segmentation

Yung-Chun Liu[1,4], Hui-Hsuan Shih[1,4], Tai-Hua Yang[2,5], Hsiao-Bai Yang[3,6], Dee-Shan Yang[7], and Yung-Nien Sun[1,4]

[1] Department of Computer Science & Information Engineering
[2] Institute of Biomedical Engineering
[3] Department of Pathology, Medical College
[4] Medical Device Innovation Center, National Cheng Kung University, Taiwan, R.O.C.
[5] Orthopedic Biomechanics Laboratory, Division of Orthopedic Research, Mayo Clinic Rochester, Rochester, Minnesota
[6] Department of Pathology
[7] Department of Orthopedic Surgery, Ton-Yen General Hospital, Taiwan, R.O.C.
ynsun@mail.ncku.edu.tw

Abstract. Measurement of pathological change in pulley tissue is an important index for trigger finger disease. However, the current measurement process is mostly based on manual estimation which is subjective and time-consuming. We hence propose an automatic method for quantitatively measuring the pathological change of pulley tissue from microscopic images. We first apply the color normalization to normalize all the acquired images. Then we use a three-stepped color segmentation process to extract the areas of diseased tissues. On the other hand, we also apply an active double thresholding scheme to segment the nuclei and extract shape features of nucleus. At last, the ratio of abnormal tissue area and the ratio of abnormal nuclei are calculated as the indices for the evaluation of trigger finger disease. The result showed good correlation between the expert judgments and the measured parameters.

Keywords: Trigger finger, pulley, segmentation, color normalization, double thresholding.

1 Introduction

Trigger finger is a common acquired condition in which the sheath for the flexor tendon of a finger thickens and narrows such that the flexor tendon cannot glide through it smoothly. This may cause pain, intermittent snapping ("triggering") or actual locking (in flexion or extension) at the affected finger [1]. The pathological change in the flexor sheath is fibrocartilaginous metaplasia (or chondroid metaplasia) of its "A1" pulley [2]. In normal pulley, there is the dense regular connective tissue that composed of collagenous fibers in compact and parallel bundles. Generally, pathohistological tissue specimen of collagenous fibers appears eosinophilic and pink

J.-S. Pan, S.-M. Chen, N.T. Nguyen (Eds.): ACIIDS 2012, Part III, LNAI 7198, pp. 476–485, 2012.

in color under Hematoxylin and Eosin (H&E) stain. Rows of modified fibroblasts with long rod-like nuclei between the collagenous bundles in longitudinal section can be observed. However, the pulley of trigger finger demonstrates fibrocartilaginous metaplasia (or chondroid metaplasia) which characterized by the presence of chondrocytes (cartilage cells) that have round nuclei and increase of extracellular cartilage matrix. This cartilage matrix contains three kinds of glycosaminoglycans: hyaluronic acid, chondroitin sulfate and keratan sulfate. Because large numbers of sulfate groups, proteoglycans, were stained by basic dyes and hematoxylin, it appeared basophilia (blue or purple color) in stained sections [3]. In this paper, we design and develop a new system for automatic microscopic tissue evaluation. The system provides two main functions, one is to segment the normal and diseased tissue areas and the other is to segment the nuclei. The ratios of abnormal areas and nuclei are then computed and used for pulley disease evaluation.

In [4], Tabesh proposed an automatic prostate cancer classification system to analyze the microscopy of the prostate cancer tissues with color features in R, G and B channels of the acquired images. However, as the acquired images are with non-uniform illumination, their simple thresholding method is not directly applicable in our case. To solve the illumination problem, we adopt the color adjusting scheme in [5] to normalize color distribution before the segmentation procedures. In [6], Wu proposed a live cell image segmentation method to directly segment the cell regions in gray level. Nevertheless, in our case, the pink areas which represent the normal tissues and the purple areas which represent the diseased tissues show very close values in gray level in the acquired images. Thus, instead of using the gray level, we have to apply the hue component of HSI color space to distinguish these two areas. After the HSI transformation, the pink normal tissues appear darker in hue component, while the purple diseased tissues are brighter in hue. Therefore, we can extract the abnormal areas in hue component more accurately than in gray level. The canny edge detector [7] is also a popular way to detect the border of cells. However, in our case, the canny operator detects not only the borders of nuclei but the borders of dark blue and noise areas. As we are only interested in the borders of nuclei, too many redundant edges detected by Canny make the application of detector impractical. In this study, we apply the active double thresholding method to segment the nuclei straightly to avoid catching the border of noise.

Our proposed automatic method quantitatively measures the pathological change of pulley tissue from microscopic images. The system contains four parts, which are color normalization, color segmentation, nuclei classification, and parameter estimation, as shown in Fig. 1. In the first part, we apply the color normalization to reduce the influence of non-uniform distribution in color and illumination of the captured images. In the second part, we transform the normalized image into HSI color space, and then use a three-stepped color segmentation process to segment the pink collagenous tissue, dark blue abnormal tissue and background areas. In the third part, we apply an active double thresholding to segment the nuclei and extract their shape features which are then used to classify the types of nuclei. In the last part, two indices which are the ratio of abnormal tissue area and the ratio of abnormal nuclei are used for the evaluation of trigger finger disease. Details of the method are described in the following sections.

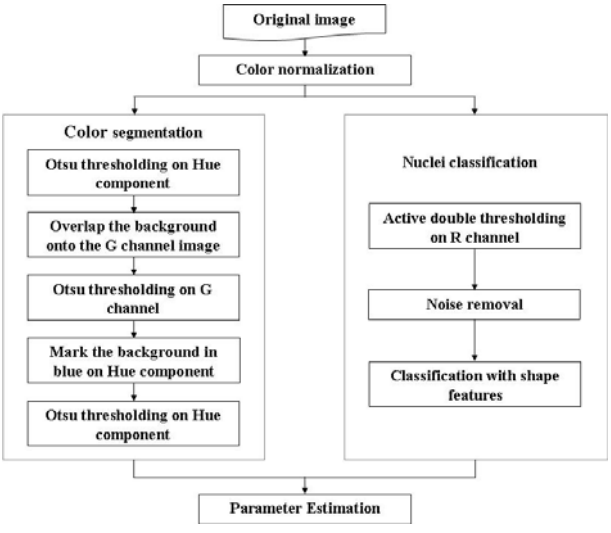


Fig. 1. Flowchart of the proposed system

2 Materials

The microscopic images of specimens in this study were provided by the laboratories in National Cheng Kung University Hospital and in Ton Yen General Hospital. The pathological pulley tissue specimens were obtained from the patients who were diagnosed with trigger finger disease by orthopedists Yang, D.S. and Yang, T.H. clinically. For pathological examination, all of the specimens followed the procedures as fixation in formalin, procession in graded alcohols and xylene, embedding in parrafin, cutting of sections with a microtome, and stained with hematoxyline-eosin (H&E). The microtome was preset for a 5-um in thickness.

In these specimens, the normal pulley showed a dense regular fibrotic tissue. The collagenous fibers were arranged in compact, parallel bundles. Between the bundles were rows of modified fibroblasts with elongated spindle-shape nuclei. The pathologic pulley tissue presented fibrocartilage metaplasia. It was composed of irregular connective tissue with fibrocartilaginous metaplasia (or chondroid metaplasia).

In the H&E stained slides, the nuclei were dark blue in color and the collagenous fibers were pink in color. The fibrocartilaginous metaplastic (or chondroid metaplastic) tissue demonstrated more chondromyxoid materials (including hyaluronic acid, chondroitin sulfate and proteoglycan) and showed blue or purple in color. Furthermore, nuclei of cartilage-like cells were also with round shape.

The prepared slides were first observed and graded according to the severity of myxoid metaplasia by pathologist Yang, H.B. under a light microscope (Olympus, BX50). On the other hand, these specimens were also analyzed by the proposed system based on the above-mentioned color and shape features. The automatic evaluation results were then compared with the golden standard by expert judgment.

3 Methods

3.1 Color Normalization

The color normalization process is used to resolve the problem of non-uniform distribution in color and illumination of the acquired images, which are caused by different staining and imaging conditions of the microscopic slice. As shown in Fig. 2(a) and 2(c), color distributions of the two acquired images are quite different from each other.

The color normalization method is adopted and modified from Reinhard [5]. At first, we must choose some standard images (target images) with the following characteristics: the contrast ratio is high and the color of nuclei is dark blue. We then normalize the input (or source) image to the color distribution of target images.

We first transform the images from RGB color space into LMS color space by the following equation:

$$\begin{bmatrix} L \\ M \\ S \end{bmatrix} = \begin{bmatrix} 0.3811 & 0.5783 & 0.0402 \\ 0.1967 & 0.7244 & 0.0782 \\ 0.0241 & 0.1288 & 0.8444 \end{bmatrix} \begin{bmatrix} R \\ G \\ B \end{bmatrix}. \tag{1}$$

Then, we transform LMS color space into log space to reduce the impact of deflection by using (2). Thereafter, the transformed space is modified with matrices multiplication in (3) to obtain the $l\alpha\beta$ space which is later used for color normalization.

$$L' = \log L, \quad M' = \log M, \quad S' = \log S, \tag{2}$$

$$\begin{bmatrix} l \\ \alpha \\ \beta \end{bmatrix} = \begin{bmatrix} \frac{1}{\sqrt{3}} & 0 & 0 \\ 0 & \frac{1}{\sqrt{6}} & 0 \\ 0 & 0 & \frac{1}{\sqrt{2}} \end{bmatrix} \begin{bmatrix} 1 & 1 & 1 \\ 1 & 1 & -2 \\ 1 & -1 & 0 \end{bmatrix} \begin{bmatrix} L' \\ M' \\ S' \end{bmatrix}. \tag{3}$$

We then calculate the mean and the standard deviation values of l, α and β for all target images and obtain the averaged mean and averaged standard deviation which are denoted as μ_t^l, μ_t^α and μ_t^β, and σ_t^l, σ_t^α and σ_t^β, respectively. These average mean and standard deviation values are calculated once and then used for the normalization of every input image. For each input image, we have to calculate the mean and standard deviation values denoted as μ_s^l, μ_s^α and μ_s^β, and σ_s^l, σ_s^α, σ_s^β, respectively.

The normalization of an input image is performed by calculating the new color values l'', α'' and β'' for each pixel by the following equations:

$$l^* = l - \mu_s^l, \quad \alpha^* = \alpha - \mu_s^\alpha, \quad \beta^* = \beta - \mu_s^\beta. \tag{4}$$

$$l' = \frac{\sigma_t^l}{\sigma_s^l} l^*, \quad \alpha' = \frac{\sigma_t^\alpha}{\sigma_s^\alpha} \alpha^*, \quad \beta' = \frac{\sigma_t^\beta}{\sigma_s^\beta} \beta^* . \tag{5}$$

$$l'' = l' + \mu_t^l, \quad \alpha'' = \alpha' + \mu_t^\alpha, \quad \beta'' = \beta' + \mu_t^\beta . \tag{6}$$

Finally, we transform the resulting image in $l\alpha\beta$ color space back to RGB color space by using (7).

$$\begin{bmatrix} R \\ G \\ B \end{bmatrix} = \begin{bmatrix} 4.4679 & -3.5873 & 0.1193 \\ -1.2186 & 2.3809 & -0.1624 \\ 0.0497 & -0.2439 & 1.2045 \end{bmatrix} \begin{bmatrix} \exp(\frac{\sqrt{3}}{3} l'' + \frac{\sqrt{6}}{6} \alpha'' + \frac{\sqrt{2}}{2} \beta'') \\ \exp(\frac{\sqrt{3}}{3} l'' + \frac{\sqrt{6}}{6} \alpha'' - \frac{\sqrt{2}}{2} \beta'') \\ \exp(\frac{\sqrt{3}}{3} l'' - \frac{\sqrt{6}}{6} \alpha'' + 0\beta'') \end{bmatrix} . \tag{7}$$

Fig. 2(b) and 2(d) show that the normalization results of Fig. 2(a) and 2(c), respectively. The color distributions of the normalized images are close to that of the target images. All input images from different batch of specimens can be processed by this procedure for color normalization.

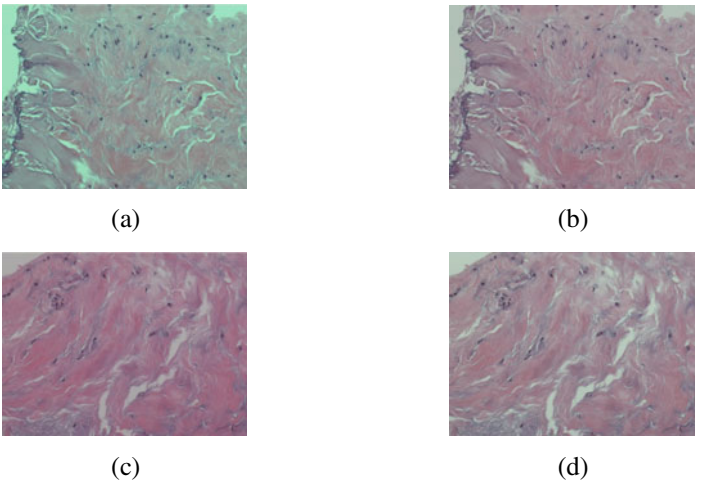


Fig. 2. Color normalization. (a) and (c) are two examples of original image; (b) and (d) are the normalized results of (a) and (c), respectively.

3.2 HSI Model Transformation and Automatic Thresholding

In color segmentation, we first transform the normalized image into the HSI color space by using (8) [8-9]. The pink part and major part of the purple areas in the normalized image are lower in hue value, and the background and some small part of the purple areas in the normalized image are with higher hue values. Fig. 3(a) shows the hue component of Fig. 2(c).

$$H = \begin{cases} \theta & if\ B \le G \\ 360 - \theta & if\ B > G \end{cases}, \quad \theta = \cos^{-1}\left\{ \frac{\frac{1}{2}\left[(R-G)+(R-B)\right]}{\left[(R-G)^2+(R-B)(G-B)\right]^{\frac{1}{2}}} \right\}. \tag{8}$$

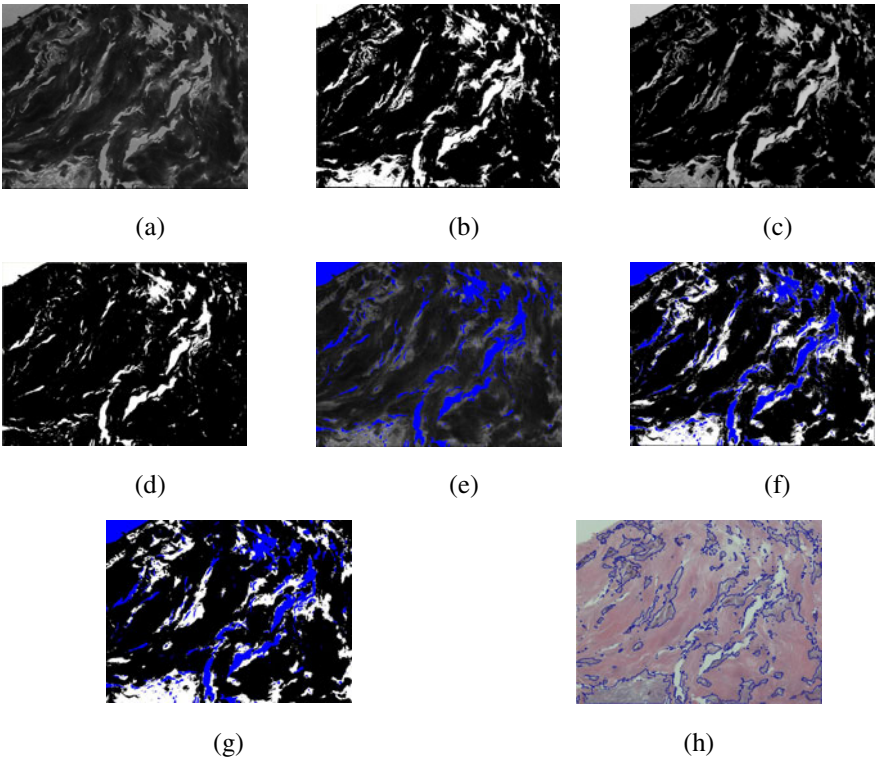


Fig. 3. Color segmentation. (a) the hue component of Fig. 2(b); (b) the Otsu threshold result of (a); (c) overlap the G Channel to the white areas of (b); (d) the result of segmentation on (c), white areas present empty background and black areas present tissue foreground; (e) blue is the empty background and the other hue component areas are tissue foreground; (f) the segmented result, blue presents background, white presents abnormal tissue, black presents normal tissue; (g) rank filtering result; (h) overlap the edges of the abnormal tissue back to the original image.

Based on the hue distribution, we apply the automatic thresholding method proposed by Otsu [10] to obtain the first binary image as shown in Fig. 3(b), which is divided into the foreground and the background roughly. The black areas represent the pink and most of the purple tissue areas as the foreground and the white areas cover some small part of the purple tissue areas and the empty background. In other words, some purple areas may be faultily classified into the background. To make the foreground include all the purple tissues, we then have to extract the small purple part from the background areas. The obtained background areas are first used as the mask to map onto the G channel of normalized image, which is shown in Fig. 3(c), and the

second Otsu thresholding on the G channel is then applied to obtain the small purple part. We then get the second binary image as in Fig. 3(d), where the white area represents the real background and the black area represents the complete foreground of pink and purple tissue areas.

After obtaining the foreground, we then have to separate the abnormal tissue from the normal tissue areas. In Fig. 3(e), we label the background areas obtained in the previous step in blue and overlap onto the original hue component image in Fig. 3(a). As mentioned before, the normal tissue areas show lower values and the abnormal areas show higher in hue, so we can use the Otsu thresholding again to divide these two areas. The segmented result is shown in Fig. 3(f), where the blue areas represent the background, the black areas represent the normal tissues and the white areas represent the abnormal tissues. As the segmentation results are fragmented in the boundaries, we apply the rank filter to remove fragmented regions. We calculate the pixel numbers of each color in Fig. 3(f) with a 9 x 9 mask, and then assign the color with the biggest counting to the central pixel of mask; the result is shown in Fig. 3(g). Fig. 3(h) shows the edges of abnormal tissues mapped onto the normalized image.

3.3 Active Double Thresholding and Features of Nuclei

Another characteristic to evaluate the level of pathological change is the ratio of round nuclei which belongs to the abnormal cells. We can use this ratio, instead of the area ratio, to characterize the tissue condition when the staining colors are faded out or specimens are after long preserving time.

After color normalization, we find that the R channel of the normalized image is more suitable for nuclei segmentation due to its high contrast of nuclei as in Fig. 4(b). Therefore, we use double thresholding scheme [11] to segment the nucleus areas. The intensity of nuclei is almost the darkest of the whole R channel image. As the intensity distributions of images are different, we thus apply an active thresholding scheme to satisfy all different images. First, for each input R channel image, we take the average of the ten lowest intensity values as the lowest intensity value of the image. Second, we add two empirical values 30 and 45 to this lowest value, and use them as the two values for double thresholding. The lower threshold value is used as the seed and the higher threshold value is the restriction of region growing. After we apply the double thresholding scheme, the white areas of the resulting image represent the segmentation of nuclei, and the segmentation result of nuclei is shown in Fig. 4(c).

Now we can classify the segmented nuclei into three categories according to their shapes. The normal nuclei are usually long rod-like, and the abnormal nuclei are usually round in shape. However, the connected area with multi-nuclei is irregular in shape and regarded as abnormal because only the abnormal nuclei will grow and connect each other into a cluster.

To classify these nuclei, we then calculate the area size, the circularity index and the maximum and the minimum distances between centroid and boundary points for each nucleus area. We then classify the nucleus is a normal rod-like if the circularity index is less than 0.95, the ratio of maximum to minimum distance is greater than 3, and the area is less than 2000 pixels. All other areas are then classified as abnormal

nuclei. In addition, we also define the area as a single abnormal round nucleus if the area is less than 2000 pixels.

After defining all the single abnormal round nuclei, we then calculate the average area of these nuclei. The average area is then used to calculate how many nuclei in a connected multi-nuclei area. Fig. 4(d) shows the classification results where red presents the normal nuclei, green presents the abnormal nuclei.

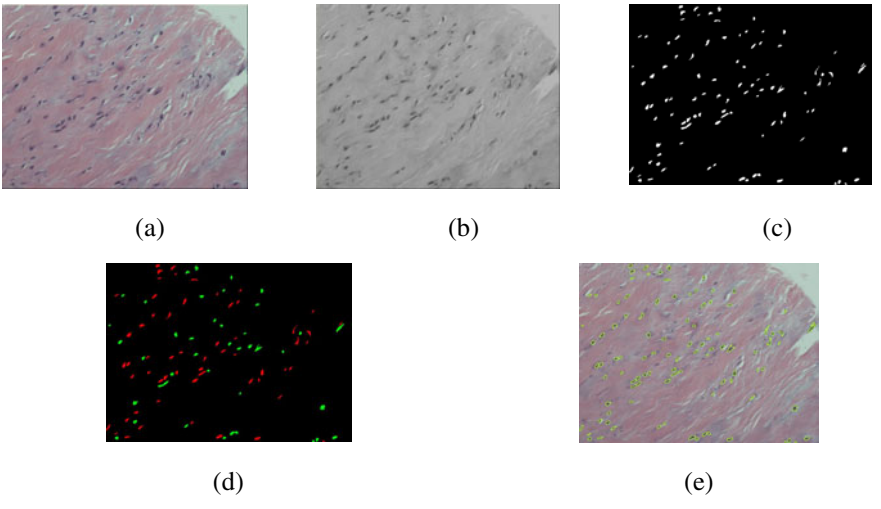


Fig. 4. Nuclei classification. (a) the normalized image of H level; (b) the R channel of (a); (c) the result after double thresholding; (d) the classification result, red presents the normal nuclei, green presents the abnormal nuclei; (e) overlap the nucleus edges onto the original image.

4 Experimental Results

In this paper, all the slides used in the experiment were acquired from the patients with trigger finger disease. The pathologist marked the lesion area for each slide, and then 49 images with the size of 2560x1920 were acquired by an auto-focusing system developed by our laboratory [12]. Next, the pathologist picked up 10 images suitable for further analysis, and ignored the rest images which were with high proportion of background or microvascular.

Table 1 lists the results obtained by the proposed color segmentation system. The normal area values represent the sum of pink areas in 10 images (of 1 slide), and the abnormal area values refer to the sum of blue areas. The abnormal area ratios show good correspondence with the expert judgment which is leveled as L, M, and H three stages. The area ratios of the three stages are significantly different as shown in Table 1. Similarly, the ratio of abnormal nuclei also shows similar results in Table 2.

For an image of 2560x1920 pixels, the average computational time of color normalization, color segmentation and nuclei classification was about 5, 12 and 10 seconds, respectively. The quantitative measurement of the two parameters could be achieved automatically and precisely by the proposed system.

Table 1. The ratio of abnormal tissue area. The normal area values represent the sum of pink areas in 10 images (of 1 slide), and the abnormal area values refer to the sum of blue areas.

Slide	Normal (pixel^2)	Abnormal (pixel^2)	Ratio
L-1	37621769	5056548	0.118
L-2	31637303	3019166	0.087
L-3	35238638	3846726	0.098
L-4	36562652	5825035	0.137
L-5	34776286	3923570	0.101
Average	35167330	4334209	0.108
M-1	36261311	7111568	0.164
M-2	34606943	7948805	0.188
M-3	36177833	7408337	0.170
M-4	30651906	7161168	0.189
M-5	35036680	8158244	0.189
Average	34546935	7557624	0.180
H-1	32011756	12057831	0.274
H-2	30452340	10354532	0.253
H-3	32757733	9434513	0.224
H-4	32319598	9777260	0.232
H-5	34071470	9897472	0.225
Average	32322579	10304322	0.242

Table 2. The ratio of abnormal nuclei. The normal values represent the sum of normal nuclei in 10 images (of 1 slide), and the abnormal values refer to the sum of normal nuclei.

Slide	Normal	Abnormal	Ratio
L-1	520	636	0.550
L-2	310	205	0.398
L-3	404	477	0.541
L-4	862	1165	0.575
L-5	292	314	0.518
Average	477.6	559.4	0.516
M-1	382	687	0.643
M-2	318	591	0.650
M-3	192	360	0.652
M-4	131	187	0.588
M-5	308	652	0.679
Average	266.2	495.4	0.642
H-1	271	660	0.709
H-2	481	1088	0.693
H-3	244	529	0.684
H-4	117	385	0.767
H-5	289	1098	0.792
Average	280.4	752.0	0.729

5 Conclusion

In this paper, we develop an automatic system to segment the diseased areas and the abnormal nuclei of pulley tissue on the microscopic image in trigger finger disease. We first apply the color adjusting scheme to normalize color distribution in the acquired images. Then we use a three-stepped color segmentation process to extract the diseased tissue areas. In addition, we also apply an active double thresholding to segment the nuclei. At last, the ratio of abnormal tissue area and the ratio of abnormal nuclei are calculated as the indices for the evaluation of trigger finger disease. Experimental results showed good correlation between the expert judgments and the measured parameters for the evaluation of pathological change in pulley tissue. The proposed system provides a reliable and automatic way to obtain pathological parameters instead of manual evaluation which is with the intra- or inter-operator variation problems. In the future, we will apply more cases for clinical validation and explore its applicability in clinic.

Acknowledgments. The authors would like to express their appreciation for the grant under contract NSC 99-2627-B-006-010 from the National Science Council, Taiwan, R.O.C.. This work also utilized the shared facilities supported by the Medical Device Innovation Center, National Cheng Kung University, Tainan, Taiwan, R.O.C.

References

1. Drossos, K., Remmelink, M., Nagy, N., de Maertelaer, V., Pasteels, J.L., Schuind, F.: Correlations between clinical presentations of adult trigger digits and histologic aspects of the A1 pulley. J. Hand Surg. 34(8), 1429–1435 (2009)
2. Sampson, S.P., Badalamente, M.A., Hurst, L.C., Seidman, J.: Pathobiology of the human A1 pulley in trigger finger. J. Hand Surg. 16(4), 714–721 (1991)
3. Sbernardori, M.C., Bandiera, P.: Histopathology of the A1 pulley in adult trigger fingers. J. Hand Surg. 32(5), 556–559 (2007)
4. Tabesh, A., Kumar, V.P., Pang, H.Y., Verbel, D., Kotsianti, A., Teverovskiy, M., Saidi, O.: Automated prostate cancer diagnosis and gleason grading of tissue microarrays. In: Proc. SPIE, vol. 5747, pp. 58–70 (2005)
5. Reinhard, E., Adhikhmin, M., Gooch, B., Shirley, P.: Color transfer between images. IEEE Computer Graphics and Applications 21(5), 34–41 (2001)
6. Wu, K., Gauthier, D., Levine, M.D.: Live cell image segmentation. IEEE Trans. Biomed. Eng. 42(1), 1–12 (1995)
7. Canny, J.: A Computational Approach to Edge Detection. IEEE Transactions on Pattern Analysis and Machine Intelligence 8(6), 679–698 (1986)
8. Gonzalez, R.C., Woods, R.E.: Digital Image Processing, 3rd edn. Prentice Hall, Upper Saddle River (2008)
9. Cheng, H.D., Jiang, X.H., Sun, Y., Wang, J.L.: Color image segmentation: advances and prospects. Pattern Recognition 34(12), 2259–2281 (2001)
10. Otsu, N.: A threshold selection method from gray-level histograms. IEEE Trans. Sys. Man. Cyber. 9, 62–66 (1979)
11. Chen, Q., Sun, Q.S., Heng, P.A., Xia, D.S.: A double-threshold image binarization method based on edge detector. Pattern Recognition 41(4), 1254–1267 (2008)
12. Liu, Y.C., Hsu, F.Y., Chen, H.C., Wang, Y.Y., Sun, Y.N.: A coarse to fine auto-focusing algorithm for microscope image. In: International Conference on System Science and Engineering (2011)

Quantitative Measurement of Nerve Cells and Myelin Sheaths from Microscopic Images via Two-Staged Segmentation

Yung-Chun Liu[1,3], Chih-Kai Chen[1,3], Hsin-Chen Chen[1,3], Syu-Huai Hong[1,3], Cheng-Chang Yang[2,3], I-Ming Jou[3,4], and Yung-Nien Sun[1,3]

[1] Department of Computer Science and Information Engineering
[2] Institute of Basic Medical Sciences
[3] Medical Device Innovation Center, National Cheng Kung University
[4] Department of Orthopedics, National Cheng Kung University Hospital, Taiwan, R.O.C.
ynsun@mail.ncku.edu.tw

Abstract. Cell morphology measurement is important in evaluating the injury level of nervous system. However, current measurement process is mostly achieved by manual estimation which is subjective and time-consuming. We hence propose a two-stage method to automatically segment axons and myelin sheaths from microscopic images for measuring the cell morphology quantitatively. First, an automatic thresholding method is used to obtain axon candidates and then geometric and image properties are used to assure the axon regions. Second, the outer contour of myelin sheath is segmented using the active contour model with the obtained axon contour as the initial solution. Then, the desired morphological parameters can be readily measured. In the experiments, we used seven nerve images for accuracy validation and achieved very small contour distance errors (less than 0.5 μm with nerve diameter around 8 μm in average). Overall, the proposed method is found efficient and useful in nerve parameter evaluation.

Keywords: Axon, myelin sheath, nerve segmentation, auto-thresholding, active contour model.

1 Introduction

The nerve evaluation process based on cell morphology measurement which provides cell structural information is an essential step for the biomedical research of nerve injury or recovery. However, manually measurement which is laborious, time-consuming and operator-dependent remains the most common approach nowadays. Unfortunately, it becomes an obstacle when the inspection of a large amount of specimens is needed. Therefore, several studies on image processing have been dedicated in developing the automatic methods to segment nerve cells on microscopic images for quantitatively measuring the cell morphology.

The canny edge detector [1-2] and watershed algorithm [3-4] are efficient methods for delineating cell contours. They usually work well on images with good intensity

J.-S. Pan, S.-M. Chen, N.T. Nguyen (Eds.): ACIIDS 2012, Part III, LNAI 7198, pp. 486–494, 2012.

contrast. Unfortunately, gradient of contours for closely attached nerve cells is sometimes weak or broken and can not always meet the requirement. Therefore, the above-mentioned methods are not directly applicable to the nerve segmentation problem. Phukpattaranont [5] proposed a counting strategy of axons based on image morphology and cell property. First, a local adaptive thresholding method was applied. Second, cell knowledge and morphology information were adopted to extract the axon regions from the foreground. Then, a modified segmentation method based on watershed algorithm was used to separate the attached cells into individual ones. Finally, the number of axons was calculated. In [6], Richerson utilized the ridge detection to enhance the intensity contrast around the myelin sheath and then employed the geometric properties to separate axons on the microscopic image. They also chose the generalized Hough transformation to estimate the size and thickness of cells. However, as the shape constraints are applied, their method may obtain unsatisfactory estimation when nerve cells are not perfectly circular or elliptical.

In this paper, we propose a two-stage segmentation method for quantitatively measuring the morphology parameters of nerve cells. The two proposed stages are axon contour estimation and myelin sheath outer contour fitting. First, we use an automatic thresholding method to obtain axon candidates and then applied the geometric and image properties to assure the axon regions. Second, based on the obtained axon contour, we then use the active contour model to estimate the outer boundary of myelin sheaths. Finally, parameters of cell morphology are computed from the segmentation results. Overall, the proposed method obtains satisfactory segmentation results with little computational time. Moreover, parameters, including number and size of cell, thickness and averaged intensity of myelin sheath, can be easily achieved by the proposed method.

2 Materials

The microscopic images used in this study were provided by Dr. Jou's laboratory in National Cheng Kung University Hospital. The image acquisition process contains several steps. At first, the rats were deeply anesthetized using intraperitoneal chloral hydrate, after which they will be given an intracardiac perfusion with warm lactated Ringer's saline solution and 4% paraformaldehyde. After the perfusion, careful dissection of the sacral vertebrae, and a gross examination of the segment of the sciatic nerve, 3-5 mm proximal and distal to the injury site was removed and stored in 4% paraformaldehyde / 2% glutaraldehyde overnight. These specimens were then post-fixed in 1% osmic acid in 0.1 M phosphate buffer (pH 7.4) for 2 hours. After they had been dehydrated in graded alcohol, the specimens were prepared and embedded in an Epon-Araldite mixture. One-micrometer-thick semi-thin sections were cut on the ultramicrotome (Leica, Germany) and observed under a light microscope (Zeiss, Germany). Finally, digital microscopic images with pixel dimension 1388 by 1040 were saved for further analysis.

3 Methods

The proposed system consists of two stages which are the axon contour estimation and myelin sheath outer contour fitting, as shown in Fig. 1. The details are described in the following sections.

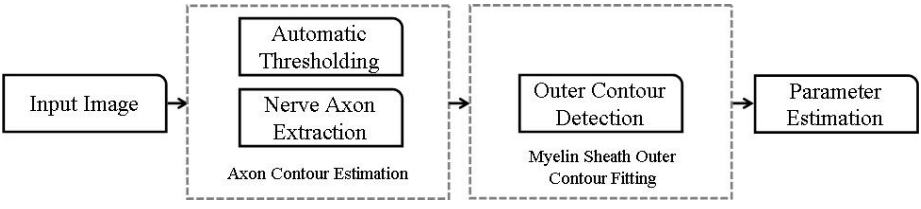


Fig. 1. Flowchart of the proposed system

3.1 Axon Contour Estimation

Automatic Thresholding. Two intensity properties of nerve microscopic image are used to segment the nerve cells. First, the axons of nerve cells are with similar intensities. Second, the intensity contrast of axon with respect to its myelin sheath is large. Based on these observations, we hence adopt the Otsu's method [2][7] to segment the axons.

Otsu's thresholding method is an automatic technique performing histogram shape-based image thresholding. This method assumes that the pixels of an image could be separated into two classes (e.g. foreground and background). It calculates the optimum threshold separating the two classes, so that the inter-class variance is maximal. Here we denote the number of pixels at i-th gray level as n_i and the total number of pixels as $N = n_0 + n_1 + \cdots + n_{L-1}$, where L is the total number of gray levels of the given image. The probability of occurrence of gray level i is defined as:

$$p_i = \frac{n_i}{N}. \tag{1}$$

Suppose that we participate the pixels into two classes C_0 and C_1 (background and foreground) by a threshold at level k; C_0 contains the pixels with levels form 0 to k and C_1 contains the pixels with levels form k+1 to *L-1*. The probabilities of the two classes and the class mean levels are given by the following equations:

$$\omega_0 = \sum_{i=0}^{k} p_i \ , \ \ \omega_1 = \sum_{i=k+1}^{L-1} p_i \ , \tag{2}$$

$$\mu_0 = \sum_{i=0}^{k} (\frac{ip_i}{\omega_0}), \ \mu_1 = \sum_{i=k+1}^{L-1} (\frac{ip_i}{\omega_1}). \tag{3}$$

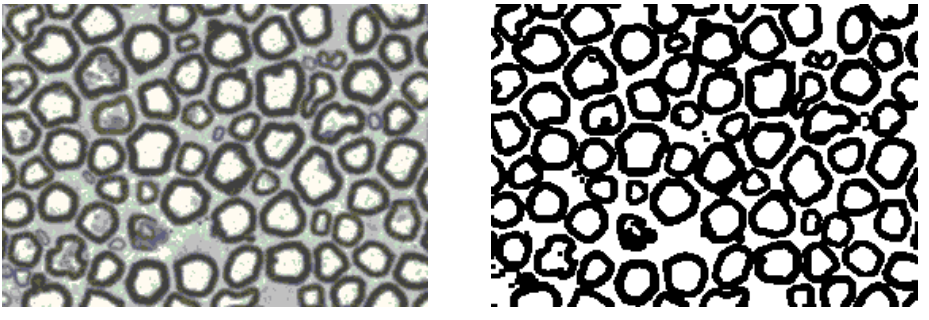
Now let σ_B^2 and σ_W^2 be the between-class variance and the within-class variance, respectively. The optimal threshold k^* can be obtained by the following equation:

$$k^* = Arg\{\max_{0 \le k \le L-1} \sigma_B^2(k)\}, \tag{4}$$

where

$$\sigma_B^2 = \omega_0\omega_1(\mu_0 - \mu_1)^2, \ \sigma_W^2 = \omega_0{\mu_0}^2 + \omega_1{\mu_1}^2. \tag{5}$$

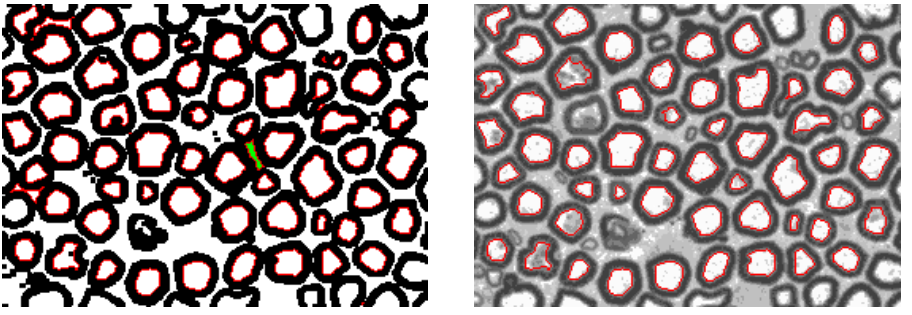
After the given image is processed by the Otsu's method and some morphological operations, we can obtain acceptable segmentation results of nerve axons (in white) and myelin sheaths (in black) efficiently, as shown in Fig. 2.



(a) (b)

Fig. 2. Otsu's thresholding segmentation. (a) the original image; (b) the segmentation result.

Nerve Axon Extraction Based on Geometry and Image Features. As the axons are supposed to be with certain geometry properties which include size, circularity index, and area ratio of axon to minimal bounding rectangle. The size is first employed to filter out non-axon regions. However, the result still contains non-axon regions, for instance, the green region indicated in Fig. 3(a). We then compute the circularity index and the area ratio of axon to minimal bounding rectangle for each region and use these two indices to filter out non-axon regions. The threshold values of these two indices are determined empirically. Moreover, we observe that the average gradient between axon and myelin sheath is usually much larger than the outer contour of myelin sheath. This intensity property can also be used to remove non-axon contours. The resulting contours after these filtering operations are shown in Fig. 3(b).



(a) (b)

Fig. 3. Nerve axon extraction: (a) before and (b) after the extraction

3.2 Myelin Sheath Outer Contour Fitting

In the previous step, we automatically extract the axons from the given image and obtain the axon contours of each nerve. The axon contour is also the inner contour of the corresponding myelin sheath. Based on the inner contour, we design a contour detection strategy via active contour model to estimate the outer contour of the corresponding myelin sheath. Let all the points of the *i*-th axon contour are denoted as $\mathbf{p}_i$ and m_i is the number of $\mathbf{p}_i$. The outer contour of the corresponding myelin sheath is then estimated by minimizing the following energy function *E*:

$$E(\mathbf{P}_i) = \sum_{j=1}^{m_i} (\alpha E_{smooth} + \beta E_{edge} + (1-\alpha-\beta) E_{internsity}), \tag{6}$$

where E_{smooth} is the smooth energy, E_{edge} is the edge energy, and $E_{internsity}$ is the intensity energy. α and β are the weighting factors indicating the trade-off between these energies. In our implementation, α and β were set to 0.3 in (6).

The smooth energy which is used to keep the shape of deformable contour model as smooth as possible is given as:

$$E_{smooth} = \mathbf{p}_i(j+1) + \mathbf{p}_i(j-1) - 2\mathbf{p}_i(j), \tag{7}$$

where $\mathbf{p}_i(j)$ is the j-th contour point of the deformable contour model $\mathbf{p}_i$, and (7) indicates the local shape curvature at the *j*-th point of the deformable contour model.

Next, the edge energy that is used to capture the true outer boundary of myelin sheath at the transition from dark to bright region (i.e., from inside myelin sheath region toward outside) is designed as:

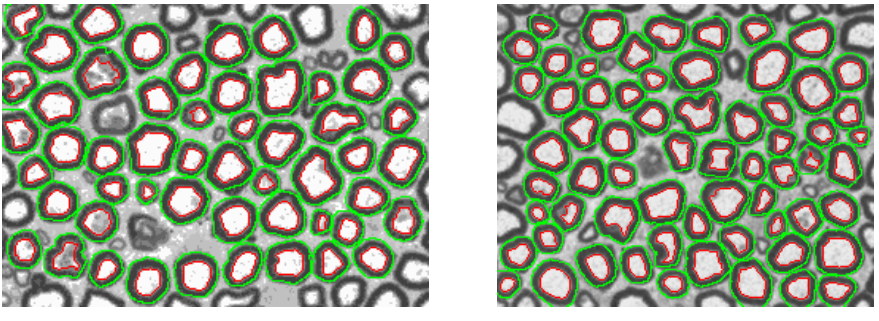
$$E_{edge} = \frac{1}{3}\sum_{a=1}^{3}[I(\mathbf{P}_i(j) + a\mathbf{n}(\mathbf{P}_i(j))) - I(\mathbf{P}_i(j) - a\mathbf{n}(\mathbf{P}_i(j)))], \quad (8)$$

where $\mathbf{n}(\mathbf{p}_i(j))$ represents the inward-pointing normal of the j-th point of the contour model. The energy is form as the intensity difference between three inner and three outer points along the normal direction.

The intensity energy, which maintains low intensity inside the myelin sheath region of the deformable contour model, is formulated as:

$$E_{intensity} = \frac{1}{3}\sum_{a=1}^{3} I(\mathbf{P}_i(j) + a\mathbf{n}(\mathbf{P}_i(j))). \quad (9)$$

This energy term is formed by the intensity sum of three pixels along the inward normal direction. And (6) is then minimized by iteratively adjusting the positions of points on the axon contour to fit the true outer boundary of myelin sheath. After the contour deformation strategy is adopted, the outer contour of myelin sheath of each nerve cell can be obtained. Finally, the structure of each nerve cell can be identified as shown in Fig. 4 (red one indicates each axon contour and green one indicates the corresponding myelin outer contour).



(a) (b)

Fig. 4. Contour fitting: (a) and (b) show the relationship of contours of each nerve on two different images (red one indicates each axon contour and green one indicates the corresponding myelin outer contour)

4 Experimental Results

In this section, the proposed method was validated with three experiments which include visual, quantitative and computational performance evaluations. For quantitative analysis, the automatic results were compared to the manual results of an expert which serve as the ground truth. The validation work was performed on a desktop PC with 3.0GHz Intel Core 2 Duo E8400 processor, Windows XP Professional, and 3GB memory.

In the visual evaluation, Fig. 5 shows the segmentation results of outer contour of myelin sheath, in which ground truth and automatic results were both superimposed onto the given image. Moreover, the average number of undetected axons with complete contour on the original image was less than 4% of the total number of axons with complete contour.

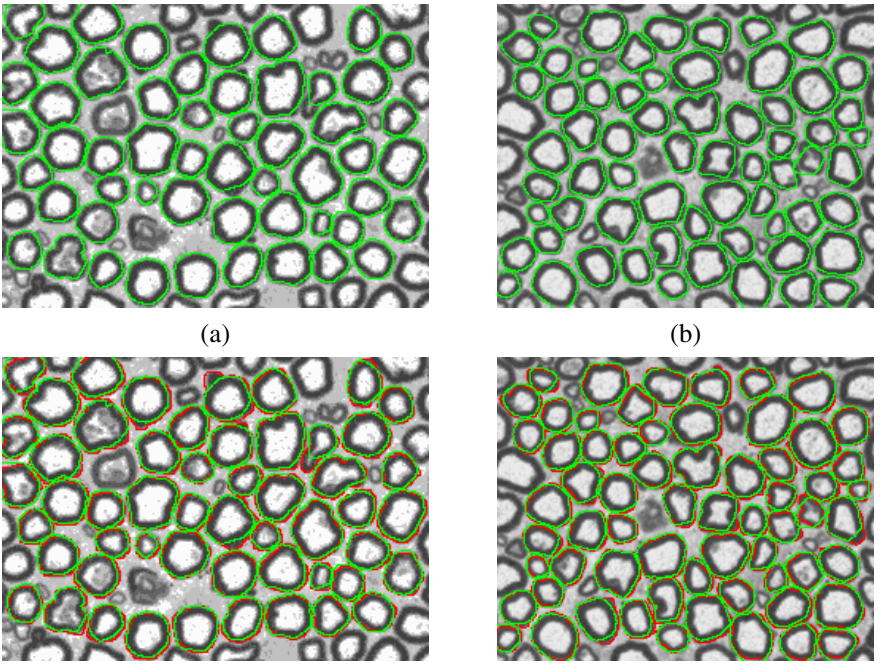


Fig. 5. Segmentation results: (a) and (b) are the automatic results of testing images; (c) and (d) show the comparison between automatic results (in green) and ground truth (in red)

As to the quantitative analysis, the comparison between the automatic results and ground truth was performed based on two difference measures of the two contours including the mean error (ME) and root mean square error (RMSE) [8]:

$$ME = \frac{1}{N}\sum_{i=1}^{N}\sqrt{(a_i - g_c)^2}\ , \tag{10}$$

$$RMSE = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(a_i - g_c)^2}\ , \tag{11}$$

where a_i is the i-th contour point of the automatic result, g_c is the c-th contour point of the ground truth which is closest to a_i, and N is the number of contour points of the automatic result. Generally, smaller values of ME and RMSE indicate that the automatic results are more consistent with the ground truth.

Table 1 lists the results of quantitative analysis. For each given image, the average values of ME and RMSE were less than 0.5 micrometer (μm). Compared with the average diameter (about 8 μm) of nerves, the segmentation errors in ME and RMSE were considered quite small (less than 6%). Based on the segmentation results, we can automatically measure the cell morphology parameters including the number of nerve cells, average size of nerve cells, thickness of each myelin sheath, and average intensity of each myelin sheath as listed in Table 2. These parameters have been reported practical and useful to neurology researches in evaluating the injury level of nerve cells.

In addition, we also evaluated the computational performance of the proposed system. The average computational time of a given image (1388 by 1040) was less than 70 seconds, which is quite small compared with the one of manual estimation. Overall, our proposed method can automatically achieve satisfactory segmentation and parameter estimation without suffering from intra- or inter-operator variability.

Table 1. Accuracy evaluation result with ME and RMSE in (mean, standard deviation)

Image	ME (μm)	RMSE (μm)
1	(0.28, 0.09)	(0.40, 0.11)
2	(0.21, 0.06)	(0.30, 0.07)
3	(0.30, 0.08)	(0.41, 0.12)
4	(0.26, 0.09)	(0.35, 0.12)
5	(0.24, 0.10)	(0.33, 0.12)
6	(0.24, 0.06)	(0.33, 0.07)
7	(0.25, 0.07)	(0.35, 0.09)
Average	(0.25, 0.08)	(0.35, 0.10)

Table 2. Morphological parameter analysis

Image	Number of Cells	Size of Nerve Cells(μm)	Thickness of Myelin Sheaths(μm)	Intensity of Myelin Sheaths
1	56	163.70	1.61	104.78
2	53	172.05	1.57	93.22
3	39	159.00	1.61	112.61
4	43	185.99	1.63	78.35
5	58	152.66	1.55	74.69
6	52	124.85	1.49	87.96
7	51	123.84	1.53	87.05
Average	49	154.58	1.57	91.24

5 Conclusion

We have proposed an automatic segmentation method for axons and myelin sheaths of nerve cells from microscopic images. First, we used an automatic thresholding method and applied the geometric and image properties to detect the axon contours. Then a contour deformation strategy using active contour model was employed to obtain the outer contour of myelin sheath. Finally, the desired morphological parameters were estimated from the segmentation results. The experimental results showed that the proposed automatic segmentation method achieved accurate parameter estimation efficiently. Thus, the proposed method can be applied clinically in evaluating structural changes of nerve cells caused by human oppression or surgery operations.

Acknowledgments. The authors would like to express their appreciation for the grant under contract NSC 99-2627-B-006-010 from the National Science Council, Taiwan, R.O.C.. This work also utilized the shared facilities supported by the Medical Device Innovation Center, National Cheng Kung University, Tainan, Taiwan, R.O.C.

References

1. Canny, J.: A Computational Approach to Edge Detection. IEEE Transactions on Pattern Analysis and Machine Intelligence 8(6), 679–698 (1986)
2. Gonzalez, R.C., Woods, R.E.: Digital Image Processing, 3rd edn. Prentice Hall, Upper Saddle River (2008)
3. Wang, Y.Y., Sun, Y.N., Lin, C.C.K., Ju, M.S.: Nerve Cell Segmentation via Multi-Scale Gradient Watershed Hierarchies. In: 28th Annual International Conf. on IEEE Engineering in Medicine and Biology Society, pp. 4310–4313 (2006)
4. Costa, A.F., Mascarenhas, N.D., De Andrade Netto, M.L.: Cell nuclei segmentation in noisy images using morphological watersheds. In: Proc. SPIE, vol. 3164, pp. 314–324 (1997)
5. Phukpattaranont, P., Boonyaphiphat, P.: An Automatic Cell Counting Method for Microscopic Tissue Image from Breast Cancer. In: 3rd Kuala Lumpur International Conference on Biomedical Engineering, pp. 241–244 (2006)
6. Richerson, S.J., Condurache, A.P., Lohmeyer, J.A., Schultz, K., Ganske, P.: An Initial Approach to Segmentation and Analysis of Nerve Cells using Ridge Detection. In: IEEE Southwest Symposium on Image Analysis and Interpretation, pp. 113–116 (2008)
7. Otsu, N.: A threshold selection method from gray-level histograms. IEEE Trans. Sys. Man. Cyber. 9, 62–66 (1979)
8. Chen, H.C., Jou, I.M., Wang, C.K., Su, F.C., Sun, Y.N.: Registration-based segmentation with articulated model from multipostural magnetic resonance images for hand bone motion animation. Medical Physics 37, 2670–2682 (2010)

Segmentation and Visualization of Tubular Structures in Computed Tomography Angiography

Tomasz Hachaj[1] and Marek R. Ogiela[2]

[1] Pedagogical University of Krakow,
Institute of Computer Science and Computer Methods,
2 Podchorazych Ave, 30-084 Krakow, Poland
tomekhachaj@o2.pl
[2] AGH University of Science and Technology
30 Mickiewicza Ave, 30-059 Krakow, Poland
mogiela@agh.edu.pl

Abstract. The new contribution of this article is description of filtering algorithm for detecting tubular structures (veins / arteries) in three-dimensional images. An algorithm incorporate the Frangi's filtration with additional neighborhood analysis filter that eliminates local noises that often remains after that algorithm. The sensitivity of the method is steered by two algorithm's parameters that might be visualized in 3D plot. Changing of those parameters does not require recalculation of filtration results. Also the concepts of those parameters are more intuitive to the potential user then the three scalable eigenvalues - based Frangi's parameters. The whole solution was tested on real volumetric CTA data.

Keywords: Artificial intelligence, Segmentation of tubular structures, Computed tomography angiography, Hessian matrix, Brain stroke.

1 Introduction

Dynamic computed tomography perfusion (DpCT) and computed tomography angiography (CTA) are two popular medical imaging methods that provide an effective add-on to standard CT in acute stroke imaging [1]. DpCT is a neuroradiology examination that enables to evaluate total and regional blood flows in time unit while CTA is imaging modality for demonstrating vascular anatomy. Imaging of the carotid arteries is important for the evaluation of patients with ischemic stroke or Transient Ischemic Attack (TIA). CTA of the head and neck is readily available and can be part of the routine imaging of stroke patients [2]. During each CTP examination huge 3D datasets is produced for which volume visualization is crucial for further diagnosis. Although different diagnosis support systems (DSS) have unique capabilities and functionality, all provide the options of volume rendering and maximum intensity projection for image display and analysis [3]. Beside visualization some DSS combines the most commonly used three-dimensional visualization techniques with image processing methods for analysis of vascular

J.-S. Pan, S.-M. Chen, N.T. Nguyen (Eds.): ACIIDS 2012, Part III, LNAI 7198, pp. 495–503, 2012.

morphology [4]. One of the basic image processing operations performed on CTA is extraction of the vessel structure. There two main groups of vascular system segmentation methods. The first one is tracking based methods. The second group takes advantage of geometric specificity of vessels, in particular the notions of orientation and tubular shape. The tracking algorithms are often based of variations of region growing algorithms [4], [5]. The second group consists of edge detection usually by a convolution filter preceded by Gaussian smoothing [6], [7], [8], [9]. The first group is dedicated mainly for detailed extraction of continues tubular structures. Second group is capable to detect any local tubular structure but is more sensitive to noises and scanning artifacts. The ability of detecting any vascular structure (without prior specification particular one of them) is especially important if there is a need to enhance the position of examined vascular tissues during the visualization process. The proper segmentation of image may help physician to localize the arteries of interests in 3D and speed up the process proper measurement that is often performed in axial slices [10] (so the high accuracy of 3D visualization is not necessarily required).

The direct volume rendering techniques enables visualization of large 3D datasets and interaction with it in real time [11]. The ISO values acquired during CT scanning (densities of visualized tissues) are often mapped to the rgba color space by so-called transfer function (the process of mapping is called "classification" [12]). Transfer function are constructed in order to visualize some ranges of tissues densities and hide the others and to make the rendered volume look realistic (Fig. 1. A and B). Unfortunately it is often impossible to use transfer function to "classify" contrasted arteries scanned during CT angiography. It is because not only contrasted tissues have given rage of density but also they might be many tissues of that kind. This happens because in real CT data there is no "sharp borders" between tissues with different density. The thresholded volumetric data in the range of the density of contrasted blood vessels consist of not only vascular systems but also the border regions between bones and sparser tissues (Fig. 1. C and D). Because of that it is impossible to segment the arteries from the rest of volume only by manipulation of transfer function.

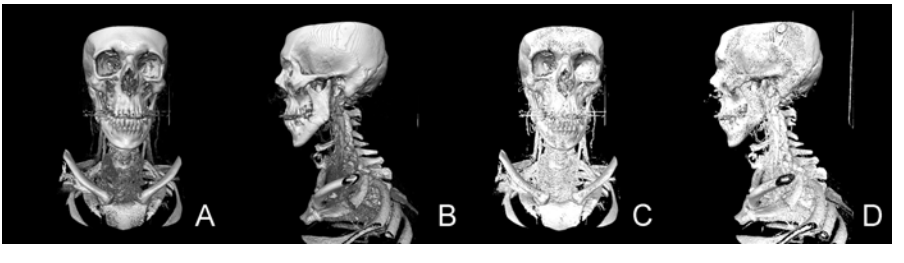


Fig. 1. Visualization of CT volume image with volume-rendering technique and proper transfer function in coronal (A) and sagittal (B) view. The same CT volume after tresholding the image in the range of the density of contrasted blood vessels in coronal (A) and sagittal (B) view.

Fig. 1. Visualization of CT volume image with volume-rendering technique and proper transfer function in coronal (A) and sagittal (B) view. The same CT volume after tresholding the image in the range of the density of contrasted blood vessels in coronal (A) and sagittal (B) view.

The new contribution of this article is description of filtering algorithm for detecting tubular structures (veins / arteries) in three-dimensional images. An algorithm incorporate the Frangi's filtration [6] with additional neighborhood analysis filter that eliminates local noises that often remains after that algorithm. The sensitivity of the method is steered by two algorithm's parameters that might be visualized in 3D plot. Changing of those parameters does not require recalculation of filtration results. Also the concepts of those parameters are more intuitive to the potential user then the three scalable eigenvalues - based Frangi's parameters. The whole solution was tested on real volumetric CTA data that makes the obtained results more reliable from those tasted on artificial phantoms.

2 Methods

A common approach to analyze the local behavior of an image I is to consider its Taylor expansion in the neighborhood of a point x_0:

$$I(x_0 + \delta x_0, s) \approx I(x_0 + \Delta x) + \delta x_0{}^T \nabla_{0,s} + \delta x_0{}^T H_{0,s} \delta x_0 \tag{1}$$

Where $\nabla_{0,s}$ is a gradient vector, $H_{0,s}$ is a hessian matrix (2) and s is a scale parameter.

$$H = \begin{pmatrix} \frac{\partial^2 I}{\partial x^2} & \frac{\partial^2 I}{\partial x \partial y} & \frac{\partial^2 I}{\partial x \partial z} \\ \frac{\partial^2 I}{\partial y \partial x} & \frac{\partial^2 I}{\partial y^2} & \frac{\partial^2 I}{\partial y \partial z} \\ \frac{\partial^2 I}{\partial z \partial x} & \frac{\partial^2 I}{\partial z \partial y} & \frac{\partial^2 I}{\partial z^2} \end{pmatrix} \tag{2}$$

The numerical gradient is often use in computer graphic for local illumination model [13]. The second partial derivatives (for example Laplacian operator) are utilized in tasks of image processing as edge detector filters. The elements of hessian matrix approximate 2nd order derivatives, and therefore encode the shape information – both a qualitative and quantitative description of how the normal to an isosurface changes [14].

Assuming that function I is continuous the hessian matrix H is symmetric. The H matrix, as a real valued and symmetric matrix, has real valued eigenvalues.

In [6] tubular structures enhancement filtering based on eigenvalues of discrete hessian matrix is proposed. Frangi et. Al. defined differentiation as a convolution with derivatives of Gaussians:

$$\frac{\partial}{\partial x} I(x,s) = s^{\gamma} I(x) * \frac{\partial}{\partial x} G(x,s) \tag{3}$$

Where γ is a parameter that define the family of normalized derivatives [6] (in our case $\gamma = 1$), $*$ is a convolution operator and G is a Gaussian kernel.

In the remainder of the paper eigenvalues will be ordered from the smallest to the largest magnitude ($|\lambda_1| \le |\lambda_2| \le |\lambda_3|$).

For each pixel of 3D image filter generates vesselnes measure defined as:

$$\nu(s) = \begin{cases} 0 \quad if \quad \lambda_1 > 0 \quad or \quad \lambda_2 > 0 \\ \left(1 - e^{\left(-\frac{{R_A}^2}{2\cdot\alpha^2}\right)}\right) \cdot \left(-e^{\left(-\frac{{R_B}^2}{2\cdot\beta^2}\right)}\right) \cdot \left(1 - e^{\left(-\frac{S^2}{2\cdot c^2}\right)}\right) \end{cases} \tag{4}$$

Where

$$R_B = \frac{|\lambda_1|}{\sqrt{|\lambda_2 \cdot \lambda_3|}} \tag{5}$$

$$R_A = \frac{|\lambda_2|}{|\lambda_3|} \tag{6}$$

$$S = \sqrt{\sum_{i=1}^{3} {\lambda_i}^2} \tag{7}$$

R_B ratio accounts for the deviation from a blob-like structure but cannot distinguish between a line- and a plate-like pattern. R_A ratio is essential for distinguishing between plate-like and line-like structures. S ratio is used to distinguish the background pixels from vessels structures (in MRA and CTA vessel structures are brighter than the background). α, β and c are thresholds that controls the sensitivity of the line filter to the measures R_A, R_B and S (in our case $\alpha = \beta = 0.5$ as suggested in [6] while c was chosen to have relatively low value $c = 3$).

The vesselness measure in Equation (4) is analyzed at different scales, s. The response of the line filter will be maximal at a scale that approximately matches the size of the vessel to detect.

$$\nu_0(\gamma) = \max_{s_{\min} \le s \le s_{\max}} \nu_0(s,\gamma) \tag{8}$$

where s_{min} and s_{max} are the maximum and minimum scales at which relevant structures are expected to be found (in our case $s_0 = 0.5$, $s_1 = 1$ and $s_2 = 2$).

The Frangi's method operates under constraints that there are no large continues regions with higher density then contrasted tissues. In order to satisfied this condition we performed additional tresholding of $v_0(\gamma)$ matrix giving value 0 to all cells for which corresponding tissue density is outside the range of interests.

The Frangi's filter is also sensitive for noises and scanning artifacts especially during filtering in small scales. These noises are visible in filtered volumes as short thick lines (Fig. 2. B). In order to eliminate them we proposed another filter analyzing the pixel neighborhood with the radius of 1 pixel in each direction. The filter response is the number of neighbor pixels which density is inside the given range. If the density of considered pixel is outside the given range the filter response is set to 0 (Fig. 2. A). The final response of our filter is a common part of Frangi's and neighbor filter at the given threshold values (Fig. 3. C).

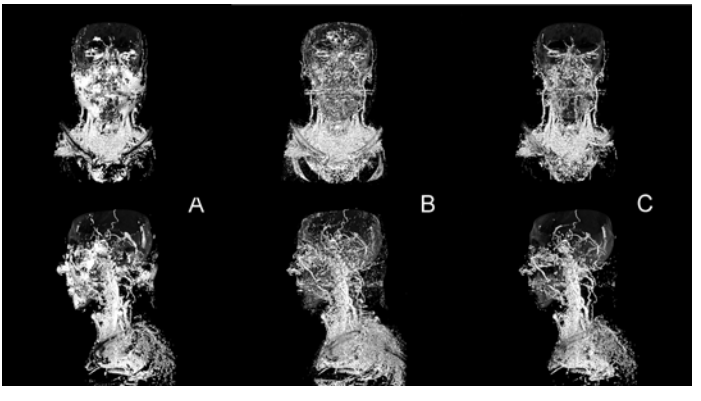


Fig. 2. Visualization of CT volume image with detected tubular structures: volume-rendering technique in coronal (top row) and sagittal (bottom row) view. The filtered regions are visible as white pixels. (A) results after neighbor filtering with threshold 15. (B) results after Frangi's filtering with threshold 0.3. (C) common part of filtered regions from (A) and (B).

Visualization of CT volume image with detected tubular structures: volume-rendering technique in coronal (top row) and sagittal (bottom row) view. The filtered regions are visible as white pixels. (A) results after neighbor filtering with threshold 15. (B) results after Frangi's filtering with threshold 0.3. (C) common part of filtered regions from (A) and (B).

3 Implementation and Results

We have implemented our filter in Matlab environment. For computation of eigenvalues of real symmetric 3x3 matrix we utilized DSYEV routine from Lapack library [15]. The direct volumetric rendering was performed with visualization

algorithm that is a part of diagnostic tool [16], [17]. The test data consisted of four CTA of carotid arteries volumes scanned by SOMATOM Sensation 10 with sizes: 512x512x415, 512x512x443, 512x512x425 and 512x512x432 voxels. Fig. 2. visualize the neighbor filtering results. The filter has capability of finding the solid regions containing voxels in given density range but it cannot differ between tubular structures and surfaces. On the other hand Frangi's filter eliminates surfaces but does not exclude short thick lines. The common part of filtered regions put together the advantages of both filtering methods. Results of detection of tubular structures in all four CTA volume datasets in coronal and sagittal view are presented in Fig. 3. Nearly all thick-lines artifact were eliminated. The elimination of the rest of them might be obtained by increasing the parameter of filtering thresholds, but it might also damage the continuity of rightly detected thick vessels. Because of that thresholds values should be carefully tuned to each case separately.

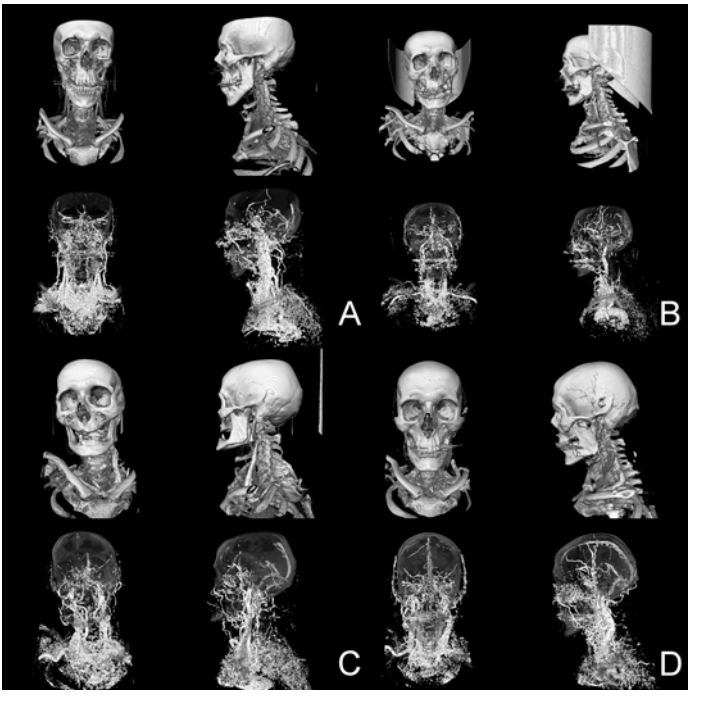


Fig. 3. Results of detection of tubular structures in four CT volume datasets – coronal and sagittal view. Results was obtained after neighbor filtering with threshold 15 and Frangi's filtering with threshold 0.3.

In Fig. 4. we present the 3D plots showing dependents of percentage of remaining voxels from a proposed filtration process as a function of vesselness coefficient and neighbors count threshold. Those plots can be use for interactive tuning of filtering

parameters. The user can easily choose the amount of threshold seeing in the same time how many voxels from the given density range has been omitted. The plots are generated for volumes from Fig. 3 and are positioned in the same order. It can be easily seen that all plots have very similar shape.

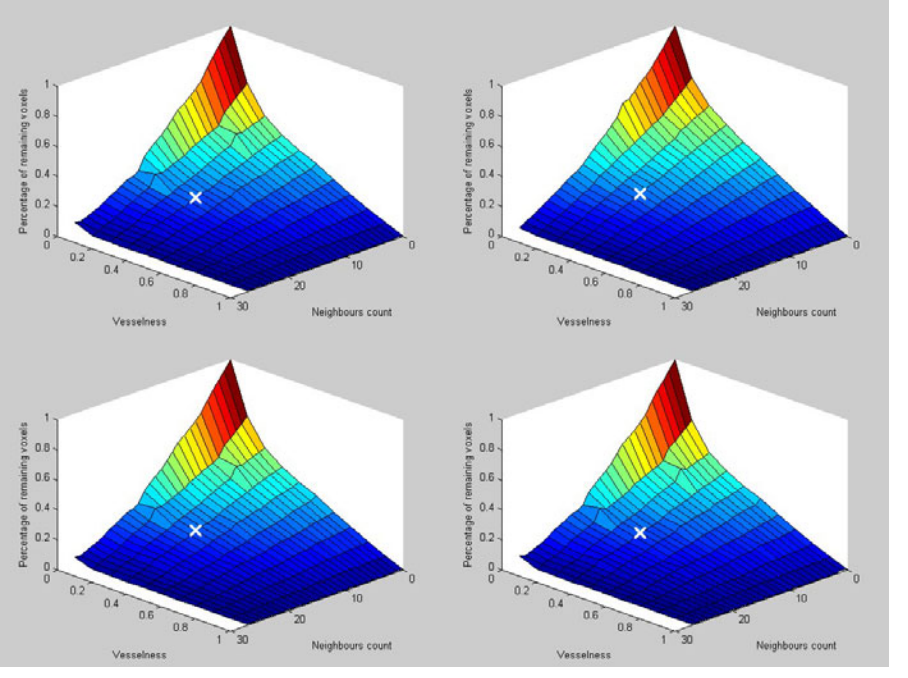


Fig. 4. The 3d plots showing dependents of percentage of remaining voxels from a proposed filtration process as a function of vesselness coefficient and neighbors count threshold for volumes from Fig. 3 (the order of plots corresponds to order of volumes in Fig. 3). The white "X" marker indicates the threshold value (vesselness = 0.3, neighbors count = 15) that was chosen in this case for filtration of tubular structures.

In Fig. 5. we present example usage of our filter for enhancing the position of contrasted vascular tissues during visualization of volumetric data (Fig. 5. bottom row). The tissues of interests are now much better visible then while they are colored only by transfer function (Fig. 5. top row). As we have proof before that kind of segmentation could not be obtained only by manipulation of transfer function because it would also affect all non-tubular tissues of the same density that are present in the visualized volume. The remaining artifact that are visible outside the regions restricted by bone structure does not disturb the process of analyzing of CTA results.

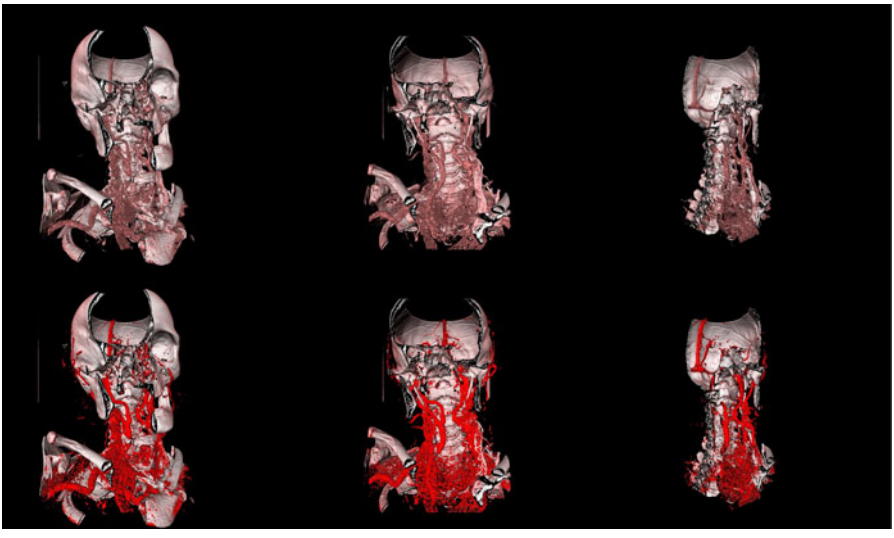


Fig. 5. Visualization of CT volume image with volume-rendering technique and proper transfer function (top row) and the same volume with enhanced tubular structures detected by proposed approach

4 Discussion

In this article we propose the new approach for detecting tubular structures (arteries) in three-dimensional images by applying proper image filtration. The filtering method proposed in this article can be use for adjusting the position of vascular tissues in volumetric and MIP visualization of 3D data. It can also be used in the preprocessing step before semantic classification of visualized symptoms [18]. Our approach can also be utilized to any image-processing problem that requires segmentation of tubular structures from 3D data. The goal of our further research will be creation of hybrid filtering method incorporating both edge detection and tracking filtering of blood vessels. The Frangi's – neighbor filtering will be initial step of the method in which all potential vascular structures will be detected. After that each tubular structure will be detailed analyzed by proper tracking method in order to improve the quality of segmentation and exclude all remaining artifacts.

Acknowledgments. This work has been supported by the Ministry of Science and Higher Education, Republic of Poland, under project number N N516 511939.

References

1. Scharf, J., Brockmann, M.A., Daffertshofer, M., Diepers, M., Neumaier-Probst, E., Weiss, C., Paschke, T., Groden, C.: Improvement of Sensitivity and Interrater Reliability to Detect Acute Stroke by Dynamic Perfusion Computed Tomography and Computed Tomography Angiography. Journal of Computer Assisted Tomography 30(1), 105–110 (2006)

2. Josephson, S.A., Bryant, S.O., Mak, H.K., Johnston, S.C., Dillon, W.P., Smith, W.S.: Evaluation of carotid stenosis using CT angiography in the initial evaluation of stroke and TIA. Neurology 63(3), 457–460 (2004)
3. Fishman, E.K., Ney, D.R., Heath, D.G., Corl, F.M., Horton, K.M., Johnson, P.T.: Volume Rendering versus Maximum Intensity Projection in CT Angiography: What Works Best, When, and Why. RadioGraphics 26, 905–922 (2006), doi: 10.1148/rg.263055186
4. Hernández-Hoyos, M., Orkisz, M., Puech, P., Mansard-Desbleds, C., Douek, P., Magnin, I.E.: Computer-assisted Analysis of Three-dimensional MR Angiograms. RadioGraphics 22, 421–436 (2002)
5. Yan, P., Kassim, A.A.: MRA Image Segmentation with Capillary Active Contour. In: Duncan, J.S., Gerig, G. (eds.) MICCAI 2005, Part I. LNCS, vol. 3749, pp. 51–58. Springer, Heidelberg (2005)
6. Frangi, A.F., Niessen, W.J., Vincken, K.L., Viergever, M.A.: Multiscale Vessel Enhancement Filtering. In: Wells, W.M., Colchester, A.C.F., Delp, S.L. (eds.) MICCAI 1998. LNCS, vol. 1496, pp. 130–137. Springer, Heidelberg (1998)
7. Olabarriaga, S.D., Breeuwer, M., Niessen, W.J.: Evaluation of Hessian-based filters to enhance the axis of coronary arteries in CT images. In: Proceedings of the 17th International Congress and Exhibition, Computer Assisted Radiology and Surgery, CARS 2003. International Congress Series, vol. 1256, pp. 1191–1196 (2003)
8. Sato, Y., Nakajima, S., Atsumi, H., Koller, T., Gerig, G., Yoshida, S., Kikinis, R.: 3D multi-scale line filter for segmentation and visualization of curvilinear structures in medical images. Medical Image Analysis 2(2), 143–168 (1998)
9. Sato, Y., Westin, C.F., Bhalerao, A., Nakajima, S., Shiraga, N., Tamura, S., Kikinis, R.: Tissue Classification Based on 3D Local Intensity Structure for Volume Rendering. IEEE Trans. on Visualization and Computer Graphics 6(2), 160–180 (2000)
10. Bartlett, E.S., Walters, T.D., Symons, S.P., Fox, A.J.: Carotid Stenosis Index Revisited With Direct CT Angiography Measurement of Carotid Arteries to Quantify Carotid Stenosis. Stroke 38, 286–291 (2007)
11. Nelson, M.: Optical models for direct volume rendering. IEEE Transactions on Visualization and Computer Graphics 1(2), 99–108 (1995)
12. Engel, K., et al.: Real-Time Volume Graphics. CRC Press (2006)
13. Phong, B.T.: Illumination for Computer Generated Pictures. Communications of the ACM 18(6), 311–317 (1975)
14. Hladuvka, J., König, A., Gröller, E.: Exploiting eigenvalues of the Hessian matrix for volume decimation. In: The 9th International Conference in Central Europe on Computer Graphics, Visualization, and Computer Vision, WSCG (2001)
15. LAPACK — Linear Algebra PACKage, `http://www.netlib.org/lapack/`
16. Hachaj, T., Ogiela, M.R.: A system for detecting and describing pathological changes using dynamic perfusion computer tomography brain maps. Computers in Biology and Medicine 41, 402–410 (2011)
17. Hachaj, T., Ogiela, M.R.: Augmented Reality Approaches in Intelligent Health Technologies and Brain Lesion Detection. In: Tjoa, A.M., Quirchmayr, G., You, I., Xu, L. (eds.) ARES 2011. LNCS, vol. 6908, pp. 135–148. Springer, Heidelberg (2011), doi:10.1007/978-3-642-23300-5_11
18. Ogiela, L., Ogiela, M.R.: Cognitive Techniques in Visual Data Interpretation. SCI, vol. 228. Springer, Heidelberg (2009)

Incorporating Hierarchical Information into the Matrix Factorization Model for Collaborative Filtering

Ali Mashhoori and Sattar Hashemi

Department of Computer Science and Engineering, Shiraz University, Shiraz, Iran
mashhuri@cse.shirazu.ac.ir, s_hashemi@shirazu.ac.ir

Abstract. Matrix factorization (MF) is one of the well-known methods in collaborative filtering to build accurate and efficient recommender systems. While in all the previous studies about MF items are considered to be of the same type, in some applications, items are divided into different groups, related to each other in a defined hierarchy (e.g. artists, albums and tracks). This paper proposes Hierarchical Matrix Factorization (HMF), a method that incorporates such relations into MF, to model the item vectors. This method is applicable in the situations that item groups form a general-to-specific hierarchy with child-to-parent (many-to-one or many-to-many) relationship between successive layers. This study evaluates the accuracy of the proposed method in comparison to basic MF on the Yahoo! Music dataset by examining three different hierarchical models. The results in all the cases demonstrate the superiority of HMF. In addition to the effectiveness of HMF in improving the prediction accuracy in the mentioned scenarios, this model is very efficient and scalable. Furthermore, it can be readily integrated with the other variations of MF.

Keywords: Collaborative Filtering, Hierarchical Matrix Factorization, Matrix Factorization, Recommender Systems.

1 Introduction

One of the main approaches to build recommender systems for e-commerce systems is collaborative filtering (CF) [1, 2] which tries to predict the preferences of a user just based on the analogies in the previous behaviors of that user and the other users without involvement in describing users or items. This advantage along with the ability of CF systems to discover unknown patterns in user behaviors, make CF to be the most applied approach in recommender systems.

Matrix factorization (MF) [3, 4] is one of the well-known methods in collaborative filtering which has received much attention recently due to its scalability, accuracy and flexibility to incorporate various types of information.

The main idea of matrix factorization is to model each user and each item by a vector consisting of a number of latent factors which are computed by decomposing the user-item rating matrix. The assumption is that user preferences and item features depend on these initially unobserved factors. More precisely the user-item rating matrix R is modeled as:

J.-S. Pan, S.-M. Chen, N.T. Nguyen (Eds.): ACIIDS 2012, Part III, LNAI 7198, pp. 504–513, 2012.

$$R \approx P^T Q \tag{1}$$

where $P \in \mathbb{R}^{K \times M}$, $Q \in \mathbb{R}^{K \times N}$ and K is the number of the latent factors in the vectors. The ith (jth) column vector of matrix P (Q), denoted by p_i (q_j) is the corresponding latent vector for the ith (jth) user (item).

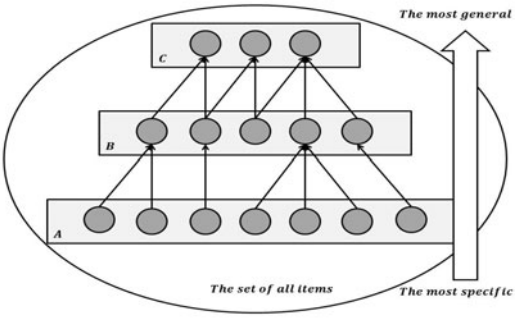


Fig. 1. The set of items is divided into 3 subsets: A, B and C, ordered from the most specific to the most general. The items in A are in a child-to-parent many-to-one relationship with the items in B. The items in B are in child-to-parent many-to-many relationship with the items in the subset C.

In many studies, researches have investigated the effect of using different matrix factorization methods on the accuracy of the predictions in CF. In real life systems, some users have a tendency to rate either high or low values and some others prefer to rate about average. Also, some items are globally liked or disliked by the users. To factor out these effects from the interaction of the vectors in P and Q biases were added to the model [5]. Temporal dynamics were incorporated into the MF models by defining the biases and the vectors in P dependent on time [6]. In [7], neighborhood model was integrated into the latent factor model. Non-negative matrix factorization (NMF) methods [8, 9] were applied to ensure that all the components of the latent vectors are positive [10, 11]. Maximum-margin matrix factorization (MMMF) [12, 13] which uses a low-norm factorization was proposed as an alternative to the standard low-rank approximation. In [14] fast maximum margin matrix factorization was proposed to solve the problem of scalability of MMMF. Kernel matrix factorization [15] was used to model nonlinear interaction between vectors corresponding to the users and the item to obtain more accurate predictions. However, in [16] it is stated that non-linear estimators are prune to overfitting.

In spite of the considerable amount of the research on MF methods, in all the previous studies, items are considered to be of the same type. However, in some situations items belong to different groups and vary in their type. In such scenarios the relation between the items of different types is a significant source of information which should be exploited to create a robust and accurate recommender system.

Motivated by the mentioned fact, in this paper we propose Hierarchical Matrix Factorization (HMF), a method that uses the defined relations between items to model

item vectors in situations that items are divided into different groups and the groups form a general-to-specific hierarchy with two or more levels. In this study, it is assumed that the items at each level are in a child-to-parent relationship with the items at the immediate layer above them. This relation can be in the form of either many-to-one or many-to-many. Figure 1 illustrates an example of this situation.

The proposed method was assessed on the Yahoo! Music dataset which is the only available dataset, known to us, containing item groups that from a general-to-specific hierarchy. The accuracy of HMF was compared to basic MF in three different cases and the results in all the cases demonstrated that HMF was successful in using the hierarchical information to obtain more accurate predictions.

Besides the effectiveness of HMF in exploiting the hierarchical relations between the items, it is very efficient and can be applied in domains with very large datasets. Furthermore, HMF is a way to model the items vectors; therefore, it can be readily integrated with other improved variations of MF.

The rest of this paper is organized as follows. Section 2 presents a formal definition of the collaborative filtering problem. Section 3 describes the proposed method. Section 4 shows the experimental results performed on the Yahoo! Music dataset and the last section summarizes our work and its contributions; also, suggests directions for future work.

2 Problem Definition

Given a set of M users and a set of N items, r_{ij} denotes the rating that user i has given to the item j. The value of r_{ij} is in the range between a minimum value V_{min} (often zero or one) and a maximum value V_{max}. The user-item rating matrix $R \in \mathbb{R}^{M \times N}$ is defined by assigning value of r_{ij} to the $(i,j)^{\text{th}}$ component of the R. In practice, the user-item matrix is often extremely huge with hundreds of millions or billions of elements; however, since each user has rated a limited number of the items, only a small subset of its components is known to us. This set of known ratings constitutes the training set $\mathcal{K}$. The challenge of collaborative filtering is to create a model $\mathcal{M}$ to estimate the value of the unknown components based on the training set $\mathcal{K}$.

3 HMF: Incorporating Hierarchical Information into MF

In the first case, let the set of items to be divided into two disjoint subsets denoted by A and B. Members of A are associated to members of B in a child-to-parent many-to-one relationship. Consequently, items in the set B group members of A into a number of clusters. Items which belong to a cluster have some common features that should be captured by a model which tries to use hierarchical information.

To model this case, a unique vector with K components is assigned to each user and each item. The prediction for the rating given by the ith user to the jth item is computed according to:

$$\hat{r}_{ij} = \begin{cases} {p_i}^T\left(q_j + q_{p(j)}\right) & if\ j \in A \\ {p_i}^T q_j & if\ j \in B \end{cases} \tag{2}$$

where p_i and q_j denote the corresponding vectors for the ith user and the jth item, respectively. Also, $p(j)$ is a function that returns the index of the parent of the jth item. In fact, items in the set B are modeled identical to basic MF while items in A are modeled by the sum of the vector associated with the item itself and the vector corresponding to its parent. $q_{p(j)}$ is the common part between all the items that have item $p(j)$ as their parent. Predicted value greater than V_{max} or less than V_{min} are set to V_{max} and V_{min}, respectively. The model learns by minimizing the squared error:

$$\underset{p^*, q^*}{min} \sum_{\mathcal{K}} (r_{ij} - \hat{r}_{ij})^2 \tag{3}$$

Stochastic gradient descent algorithm is used to minimize (3); thus, after visiting each instance in the training set, parameters are modified in the opposite direction of the gradient:

$If\ j \in A$:

$$p_i^k \leftarrow p_i^k + \alpha(q_j^k + q_{p(j)}^k)(r_{ij} - \hat{r}_{ij})$$
$$q_j^k \leftarrow q_j^k + \alpha p_i^k (r_{ij} - \hat{r}_{ij})$$
$$q_{p(j)}^k \leftarrow q_{p(j)}^k + \alpha p_i^k \left(r_{ij} - \hat{r}_{ij}\right)$$

$If\ j \in B$:

$$p_i^k \leftarrow p_i^k + \alpha q_j^k (r_{ij} - \hat{r}_{ij})$$
$$q_j^k \leftarrow q_j^k + \alpha p_i^k (r_{ij} - \hat{r}_{ij}) \tag{4}$$

where α is the learning rate and is determined by cross-validation.

By induction, this model can be extended for the cases with multiple levels of hierarchy in which each level is in a many-to-one relationship with the parent level. For example, if a third item set named C is added to the former case, and items in the set B have a many-to-one relationship with items in this new set, the estimated ratings for items are then computed according to:

$$\hat{r}_{ij} = \begin{cases} {p_i}^T\left(q_j + q_{p(j)} + q_{p(p(j))}\right) & if\ j \in A \\ {p_i}^T\left(q_j + q_{p(j)}\right) & if\ j \in B \\ {p_i}^T q_j & if\ j \in C \end{cases} \tag{5}$$

In the second case, we cover the situation in which there is a child-to-parent many-to-many relationship between the items in the set B and the items in the set C. Since in this condition each item in B might be in relation with more than one item in the set C, and all the items in the set C are at the same level of hierarchy, it does not make sense to apply the previous method to model this kind of relationship. Firstly, the effect of the parent level on the child items should be normalized in some way. Secondly, it should be considered that the effect of all the parents on the child may not be equal and some of the relations might have higher degrees of correlation.

Suppose that the jth item in the set B is in relation with L members of the set C. The rating given by the ith user to this item is computed by:

$$\hat{r}_{ij} = \begin{cases} {p_i}^T \left(q_j + \sum_{l=1}^{L} \omega_{j,p_l(j)} \, q_{p_l(j)} \right) & if \; j \in B \\ {p_i}^T q_j & if \; j \in C \end{cases} \tag{6}$$

where $p_l(j)$ returns the index of the lth parent of the jth item and $\omega_{j,p_l(j)}$ is the importance degree of the relation between the jth item and its lth parent. Importance degrees for the relations of each item in B satisfy the condition below:

$$\sum_{l=1}^{L} \omega_{j,p_l(j)} = 1 \tag{7}$$

The updating rule for ω using stochastic gradient descent is:

$$\omega_{j,p_l(j)} \leftarrow \max\{\omega_{j,p_l(j)} + \alpha\left({p_i}^T q_{p_l(j)}\right)\left(r_{ij} - \hat{r}_{ij}\right), 0\} \tag{8}$$

After updating ω for all the relations of the jth item, they should be normalized to satisfy (11).

$$\omega_{j,p_l(j)} \leftarrow \frac{\omega_{j,p_l(j)}}{\sum_{l=1}^{L} \omega_{j,p_l(j)}} \tag{9}$$

By combing the ideas in (2) and (6) the formula for calculating the estimated rating given by a user to an item at each level of a hierarchy and in every hierarchical model in which there is a general-to-specific order between levels can be derived.

4 Experimental Results and Discussions

To investigate the accuracy of the proposed method we used the Yahoo! Music dataset which was released in KDD cup 2011 competition. This dataset is a subset of the Yahoo! Music community's preferences for various musical items. It is a very large scale dataset containing over 300 million ratings given by one million users to more than 600K items. Another distinctive characteristic of this dataset is that the items in this dataset are categorized in to four different types: tracks, albums, artists, and genres and the items are linked together within a defined hierarchy. Each track belongs to an album, it is performed by an artist and it is associated with some genres. Each album has an artist and is associated with some number of genres. There is no explicit defined relation between artists and genres.

We performed three experiments with different kind of hierarchical models: 1- Two levels with many-to-one relationship between the child and the parent layers. 2- Three levels with many-to-one relationship between the layers. 3- Two levels with many-to-many relationship between child and parent layer. In each case the behavior and accuracy of basic MF was compared to hierarchical matrix factorization proposed in this paper.

It should be noted that in this study our concentration is on the modeling of item vector. Since in all the improved variations of MF which were mentioned previously item vectors are modeled the same as basic MF, we compared our model with basic MF.

Models in all the experiments were trained using the ratings given by the first 200,000 users in the training set and the ratings of the same group in the validation set (released in KDD Cup 2011) were used to test and compare the methods. Also, the number of components of the vectors in all the experiments was set to 30.

4.1 Two Levels of Hierarchy with Many-to-One Relationship between the Layers

In the first experiment we just considered the ratings which were given to the tracks and the albums. Also, only the users for whom the sum of their album ratings and track ratings was equal or greater than 10 were taken into account. In this case the items form a two level hierarchical model with child-to-parent many-to-one relationship between the tracks and the albums. We used the selected subset of ratings to train the basic MF model and HMF model. In both cases stochastic gradient descent was used for learning. The learning rate was equal to 0.0005 for basic MF and 0.0003 for HMF. Figure 2 illustrates the RMSE of both of the methods on tracks and albums after each iteration of gradient descent.

Figure 2.a depicts the RMSE of the methods on tracks. Error for both basic MF and HMF decreases significantly in the first six iterations. This stage is followed by little changes in the next four iterations and there is almost no improvement in the last three iterations. As we anticipated, the RMSE of the proposed method is lower than the basic MF in all the iterations. The best RMSE obtained by basic MF on tracks is 28.09 while this value for HMF is equal to 27.39. The superiority of HMF over basic MF is the result of the additional hierarchical information which is used to model the tracks.

Figure 2.b illustrates the RMSE of the methods on albums. Basic MF reaches about its best result at 10^{th} iteration and after that there is almost no change. The improving

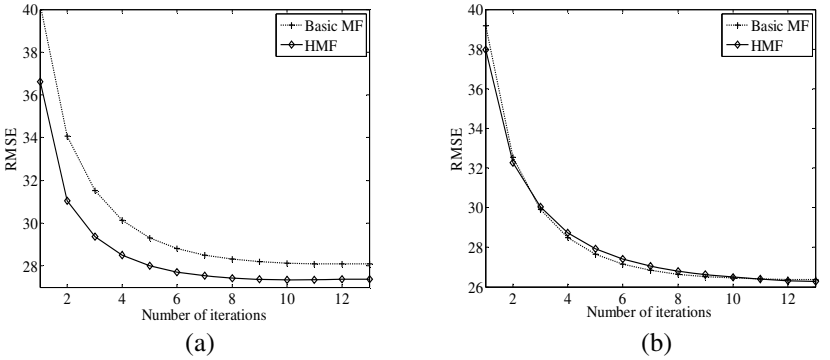


Fig. 2. Comparison between the RMSE of basic MF and HMF with two levels of hierarchy (a) RMSE on tracks: HMF outperforms basic MF due to hierarchical information which is added to model tracks. (b) RMSE on albums: The accuracy of both of the methods are equal since there is no difference in way they modeling albums.

trend of the proposed method lasts during all the iterations but there is a very slight change after the 10th iteration. Neglecting the small differences between the two methods, their accuracy on albums during all the iterations can be considered the same. This similar behavior is anticipated since in both HMF and basic MF albums are modeled in an identical way.

4.2 Three Levels of Hierarchy with Many-to-One Relationship between the Layers

In the second experiment the models were trained using the ratings given to tracks, albums and artists. In both of the training and the test phases the users for whom the sum of their track ratings, album ratings and artist ratings was less than 10 were ignored. The set of items which are considered in this case form a three layer hierarchy in which there is a many-to-one child-to-parent relationship between layers (tracks-albums, albums-artists). Basic MF model and HMF model were trained using the selected subset of rating. For both methods gradient descent algorithm was used for learning. The learning rates for basic MF and HMF were set to 0.0005 and 0.0002, respectively. Figure 3 illustrates the RMSE of the methods on tracks, albums and artists.

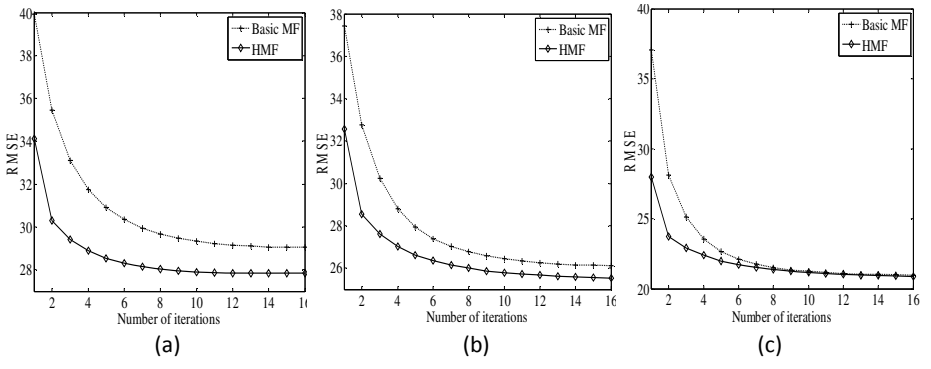


Fig. 3. Comparison between the RMSE of basic MF and HMF with three levels of hierarchy (a) RMSE on tracks: The two layer hierarchical information used to model tracks causes HMF to outperform basic MF (b) RMSE on albums: The hierarchical information used to model albums comes from one layer. The accuracy of HMF is still higher than basic MF. (c) RMSE on artists: Both of the methods model artist in the same way and their accuracy is the same in this case.

Figure 3.a compares the behavior of the methods based on their RMSE on tracks in 16 iterations. The RMSE of HMF is below basic MF in all the 16 iterations. The best result of basic MF on tracks is 29.05 while the best RMS of HMF is 27.83. Compared to the difference between accuracy of the methods on tracks in previous experiment (0.7), in this case there is a larger discrepancy between the results (1.22). The reason is that in this experiment the hierarchical information to model the tracks comes from two layers (artists and albums) while in the previous situation the information from just one layer (albums) were used to model tracks.

Figure 3.b shows the change in the RMSE of both methods on albums during 16 iterations. The behavior of both of the methods is similar to Fig 3.a. Again, the RMSE of HMF in all the iterations is below that of basic MF. The lowest RMSEs are 26.12 and 25.52 for basic MF and HMF, respectively. It should be noted that this difference (0.6) is very close to the difference of the RMSE of the two methods on tracks (0.7) in the previous experiment. The reason behind this similarity is the fact that in both of the cases the items (tracks in experiment 1 and albums here) are modeled using one layer of hierarchical information.

Figure 3.c illustrates the behavior of the methods on artists. As it was the case in the first experiment for albums, in this experiment artist are modeled identical in both basic MF and HMF. Therefore, no difference in best RMSEs of methods is observed.

4.3 Two Levels of Hierarchy with Many-to-Many Relationship between the Layers

In the last experiment we used ratings given to albums and genre to train the models. The experiment was performed on the users for whom the sum of album ratings and genre ratings was equal or greater than 10. According to the defined hierarchy in the dataset, each album is associated with one or more genres; thus, the set of these two types of items forms a two level hierarchy with many-to-many relationship between the layers. Basic MF model and HMF model were trained using the selected subset of rating. For both cases stochastic gradient descent was used for the task of learning with learning rates equal to 0.0003 and 0.0002 for basic MF and HMF, respectively. Figure 4 illustrates the RMSE of the methods on albums and genres after each iteration of gradient descent.

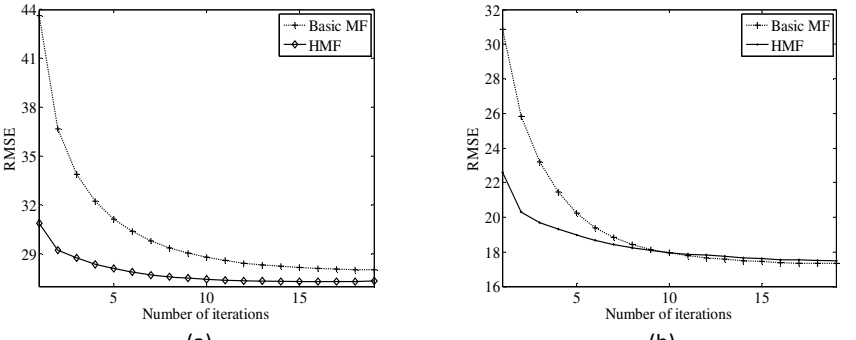


Fig. 4. Comparison between the RMSE of basic MF and HMF with two levels of hierarchy and many-to-one relationship (a) RMSE on tracks: HMF outperforms basic MF due to hierarchical information which is added to model tracks. (b) RMSE on albums: The accuracy of both of the methods are equal since there is no difference in way they modeling albums.

Figure 4.a depicts the RMSE of the methods on albums. In the 12 first iterations there is a fast and sharp drop in the RMSE of basic MF from 43.63 to 28.44. After this stage there is little improvement. The final RMSE of basic MF is 28.04. In contrast to the behavior of basic MF, the stage of sharp drop for the RMSE of HMF finishes at the second iteration. In the next eight iterations the RMSE decreases smoothly and there is almost no change afterward. The best RMSE of HMF obtained in the last iteration is 27.33. Regardless of the difference in the behavior, the RMSE of HMF in all the iterations is below the basic MF. Also, HMF reaches its stable stage around its best RMSE much earlier than basic MF. Analogous to the two previous experiments, additional hierarchical information which is used to model the albums causes HMF to outperform basic MF.

Figure 4.b illustrates the RMSE of the methods on genres. The result of this experiment is consistent with the previous experiments. Since there is no distinction between the way that basic MF and HMF model genres, no considerable difference between their accuracy on genres is observed.

5 Conclusion

In this paper, we proposed Hierarchical Matrix Factorization (HMF) for collaborative filtering to improve the accuracy of the predictions in the situations that the items are related to each other in a defined hierarchy with child-to-parent relationships between the successive layers. Two models were suggested for the both cases of many-to-one and many-to-many relationship between the child and the parent levels. By combining these two models, HMF can be applied to any hierarchical model.

The method was assessed on Yahoo! Music dataset in three experiments with a different type of hierarchy in each case. The results showed that HMF outperformed the basic matrix factorization method on all the item groups which were at the second level or the lower levels of the hierarchy. Also, the difference between the accuracy of the methods on items increased with the depth of the item group within the hierarchy since for modeling the items at the lower levels more hierarchical information is used.

In this research, our intention was to focus on the differences between the ways that HMF and other matrix factorization methods model the items. Therefore, we did not consider the effect of time or biases in the predictions. These elaborations can be readily used with HMF to obtain a more accurate model.

References

1. Goldberg, D., Nichols, D., Oki, B.M., Terry, D.: Using Collaborative Filtering to Weave an Information Tapestry. Communications of the ACM 35(12), 61–70 (1992)
2. Su, X., Khoshgoftaar, M.T.: A survey of collaborative filtering techniques. In: Advances in Artificial Intelligence, pp. 1–20 (2009)
3. Koren, Y., Bell, R., Volinsky, C.: Matrix Factorization Techniques for Recommender Systems. Computer 42(8), 30–37 (2009)

4. Takács, G., Pilászy, I., Németh, B., Tikk, D.: Investigation of various matrix factorization methods for large recommender systems. In: The 2nd KDD Workshop on Large-Scale Recommender Systems and the Netflix Prize Competition, Las Vegas, NV, USA, pp. 1–8 (2008)
5. Paterek, A.: Improving regularized singular value decomposition for collaborative filtering. In: KDD Cup and Workshop, pp. 2–5 (2007)
6. Koren, Y.: Collaborative filtering with temporal dynamics. In: The 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Paris, France, pp. 447–456 (2009)
7. Koren, Y.: Factorization meets the neighborhood: a multifaceted collaborative filtering model. In: The 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Las Vegas, USA, pp. 426–434 (2008)
8. Lee, D.D., Seung, H.S.: Learning the parts of objects by non-negative matrix factorization. Nature 401, 788–791 (1999)
9. Lee, D.D., Seung, H.S.: Algorithms for non-negative matrix factorization. In: Advances in Neural Information Processing Systems, pp. 556–562. MIT Press (2001)
10. Zhang, S., Wang, W., Ford, J., Makedon, F.: Learning from incomplete ratings using non-negative matrix factorization. In: The 6th SIAM Conference on Data Mining (SDM), Bethesda, USA, pp. 548–552 (2006)
11. Wu, M.: Collaborative Filtering via Ensembles of Matrix Factorizations. In: KDD Cup Workshop at SIGKD 2007, 13th ACM International Conference on Knowledge Discovery and Data Mining, San Jose, USA, pp. 43–47 (2007)
12. Srebro, N., Rennie, J.D.M., Jaakkola, T.S.: Maximum margin matrix factorization. In: Advances in Neural Information Processing Systems, vol. 17, pp. 1329–1336. MIT Press (2005)
13. DeCoste, D.: Collaborative prediction using ensembles of Maximum Margin Matrix Factorizations. In: The 23rd International Conference on Machine Learning, Pittsburgh, USA, pp. 249–256 (2006)
14. Rennie, J.D.M., Srebro, N.: Fast maximum margin matrix factorization for collaborative prediction. In: The 22nd International Conference on Machine learning, Bonn, Germany, pp. 713–719 (2005)
15. Rendle, S., Schmidt-Thieme, L.: Online-updating regularized kernel matrix factorization models for large-scale recommender systems. In: The 2008 ACM Conference on Recommender Systems, Lausanne, Switzerland, pp. 251–258 (2008)
16. Ott, P.: Incremental Matrix Factorization for Collaborative Filtering. Contributions to Science, Technology and Design (2008)

Optical Flow-Based Bird Tracking and Counting for Congregating Flocks

Jing Yi Tou and Chen Chuan Toh

Centre for Computing and Intelligent Systems (CCIS),
Faculty of Information and Communication Technology (FICT),
Universiti Tunku Abdul Rahman (UTAR),
Jalan Universiti, Bandar Barat, 31900 Kampar, Perak, Malaysia
toujy@utar.edu.my, agent_cc@hotmail.com

Abstract. Bird flock counting is not deeply studied yet, but there are applications of bird flock counting that would provide benefits to different parties, such as the population estimation of swiftlets for swiftlet farmers and migratory raptors for ornithologists. The methodology involved two stages: 1. segmentation phase to segment the birds out from the video and; 2. tracking phase to track the birds using optical flow and count them. The raptor flocks are used in this paper because they are bigger and easier to be filmed at congregation sites during migration. The experimental results show an average hit rate of 88.1% for video of 500 × 400 but only achieve 6.17 fps while the video of 320 × 200 had a lower hit rate of 73.3% but achieving 16.98 fps. Both are not capable of achieving real-time processing but show possibility of tracking and counting a flock of flying birds.

Keywords: Bird tracking, bird counting, raptor counting, optical flow, computer vision.

1 Introduction

Computer vision is used to create many solutions on images and videos to gain a better understanding on their contents, especially on situations that are either too tough or too tiring for the humans to work on. Birds often fly in flocks in the skies and would congregate into groups that flies together, however in order to count the number of the birds, it is a very challenging task for the human observers because the birds are hard to be tracked independently to avoid a duplicated count of the same bird that is actively moving within the group. One may count the numbers of birds from a static frame but it is hard to continue to trace the changes of numbers over some time. Such counting is beneficiary for people who are interested with the studies of certain bird population, such as ornithologist, bird-watchers, conservationist and etc.

Other than the need to study their population for scientific or conservation purposes, the estimation of numbers for certain species brings economical benefits, e.g. swiftlet population estimation. Swiftlet farming had been a popular economical

J.-S. Pan, S.-M. Chen, N.T. Nguyen (Eds.): ACIIDS 2012, Part III, LNAI 7198, pp. 514–523, 2012.

activity in South-east Asia for the recent years due to the convenience of building a swiftlet hotel in urban or sub-urban areas to attract swiftlets to breed in rather than the conventional method of harvesting swiftlet nests from natural caves [1]. The estimation of swiftlet populations at a location would be very helpful for the assessment of feasibility and sustainability for a location to conduct swiftlet farming activities.

For the ease of data collection, raptors are selected for the experiments in this paper because they are usually bigger in size and are easier to be recorded in a video sequence and remained to be visible. Raptors or birds of prey are fascinating creatures because unlike other birds, they hunt for food primarily on wing, including falcons, eagles, hawks and etc [2]. Raptors are generally counted during the migration period for the population estimation of different raptor species to help in conservation of the raptors, some important migratory points in East and South-east Asia includes Kenting National Park in Taiwan [3], Chumphon and Radar Hill in Thailand [4], Taiping and Tanjung Tuan in Peninsular Malaysia and etc [5].

The main objectives of this project are:

- To automate the tracking of each bird (raptor) individual based on a video of a flying bird flock;
- To count the number of birds (raptors) within each video frame;

For the following sections, Section 2 describes the tracking and counting algorithm developed in this paper. Section 3 describes the development tools and experimental dataset that are used for the experiments. Section 4 shares the experimental results from the conducted results and the analysis of it. Section 5 concludes the paper and describes the potential future works.

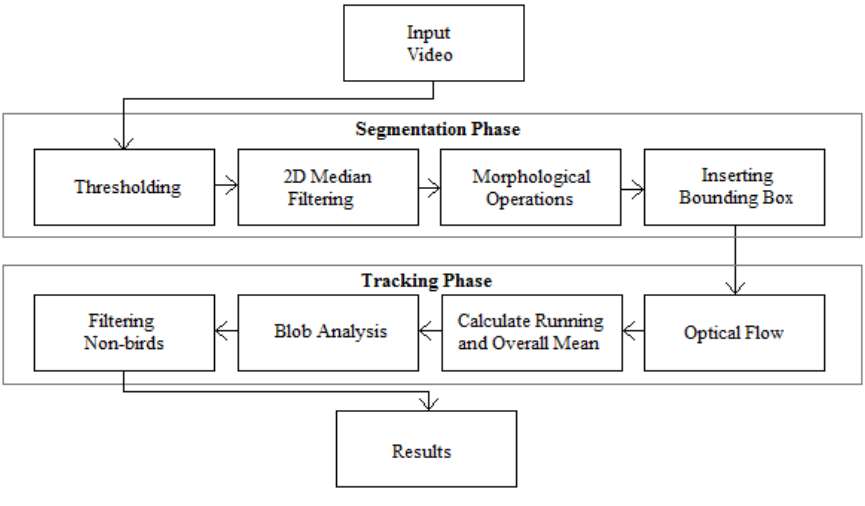


Fig. 1. Flow chart of the tracking and counting algorithm with two stages, the segmentation and tracking phases

2 Tracking and Counting

This paper uses optical flow to estimate the motion vectors in each frame of the video sequence. The methodology is mainly divided into two major phases: 1. segmentation phase and; 2. tracking phase.

By performing a threshold, filtering and morphological operations on the motion vectors, the binary feature images are produced. Rectangles are then drawn around the raptors, and a counter is used to track the number of blobs within each frame. Next, the tracking is done based on optical flow and further filter out non-birds through blob analysis. The flow of the methodology is shown in Figure 1.

2.1 Segmentation Phase

The segmentation phase is conducted to segment the target objects (birds) from the background image (sky). First, a threshold method is performed followed by a noise filter and a morphological closing and opening. Finally, the connected blobs would be bounded.

Threshold. The videos used for this paper are all against the skies with flying birds within, therefore a threshold can be applied to segment the birds from the sky by setting a threshold value [6]. The video frames will first be converted into grayscale format before the application of the threshold. The process will segment the birds into white blobs while the background is represented by black as shown in Figure 2 where the birds are segmented as white blobs along with some noise. The threshold value used in this paper is 127.

Fig. 2. Example of threshold showing the original image (left) and image after threshold (right)

Noise Filtering. After the threshold, some of the noise would also be treated as birds, however most of the blobs are usually smaller than the birds. The 3 × 3 two-dimensional median filter is used to filter the noise. This filter is a nonlinear digital filtering technique commonly used in digital image processing as it is able to remove noise whilst preserving object edges at the same time [6].

Morphological Operations. With the median filter applied, there are still traces of the small blobs and a morphological closing and opening would be applied where the morphological closing would remove gaps found within the blobs followed by the morphological opening to thin out portions between the blobs and the background [6].

The structuring element used for the morphological closing is:

0	0	0	0	1
0	0	0	1	0
0	0	1	0	0
0	1	0	0	0
1	0	0	0	0

(1)

and for the morphological opening is:

1	1
1	1

(2)

This process would remove the noises that are in the form of small blobs as shown in Figure 3.

Fig. 3. Example showing before (left) and after (right) the morphological closing and opening

Connected Component. After the processes, the connected blobs of white pixels would be considered as a single object, and a bounding box would be placed on each blob for reference.

2.2 Tracking Phase

After segmenting the birds from the videos, the next phase is to track the movements of the birds using optical flow and a final filtering of the blobs will be conducted to filter off non-birds.

Optical Flow. The optical flow is a type of vector field showing the direction and magnitude of intensity changes from one image to another. This method was first researched by James Jerome Gibson, an American psychologist in the 1940s, and uses the pattern of apparent motion of any non-rigid objects as an ingredient for motion detection and had became very popularly used for tracking moving objects in videos today [7,8]. It is also previously used for tracking birds in flight motion for the purpose of reducing bird injury [9].

The Lucas-Kanade method which was developed in 1981 by Bruce D. Lucas and Takeo Kanade is one of the most popular methods of optical flow [10]. However, the other method, Horn-Schunck method, is selected to be used in this paper. This method of optical flow was developed in 1980 by Berthold K.P. Horn and Brian G. Schunck. The advantage of this method is that it is more robust in handling image sequences which are quantized more coarsely in space and time. But due to its global approach of computing every pixel, it is more sensitive to additive noise as compared to the Lucas-Kanade method. The formula used for the Horn-Schunck method is:

$$E = \iint (I_x u + I_y v + I_t)^2 dxdy + \alpha \iint \left\{ \left(\frac{\partial u}{\partial x} \right)^2 + \left(\frac{\partial u}{\partial y} \right)^2 + \left(\frac{\partial v}{\partial x} \right)^2 + \left(\frac{\partial v}{\partial y} \right)^2 \right\} dxdy \quad (3)$$

where I_x, I_y *and* I_t are the spatiotemporal image brightness derivatives. u and v represents the horizontal and vertical optical flow relatively. $\frac{\partial u}{\partial x}$ and $\frac{\partial u}{\partial y}$ represents the spatial derivatives of optical velocity component of u and v respectively and α represents the scales the global smoothness term [11].

The system use a reference frame delay of 3 in order to measure the object's motion history without sacrificing performance. Once the optical flow of each blob has been estimated, its value of velocity can be acquired. The running and overall mean of the velocities are calculated. They are used to estimate a threshold value to define if the blob is a bird or not. If a blob in question does not exceed the threshold, it is considered to be a non-moving blob and would be therefore discarded as a bird.

Filtering Non-Birds. A blob analysis is conducted to use the parameters of the blob to determine if the blobs are birds using some simple heuristics on the parameters. The values of interest include the blob's maximum and minimum area. To further refine the analysis, the ratio between the area of the blob and the area of the bounding box are calculated. If a blob is calculated to exceed 40% of blob area within the bounding box, it is considered as a bird or else the blob will be discarded as a non-bird.

The numbers of the remaining blobs will be counted when all blobs had gone through the filtering process. The calculations will be done in every frame to determine the number of birds in the particular frame.

3 Experimental Tools and Dataset

This section described the development tools and dataset used for the experiments that are described in Section 4.

3.1 Development Tools

MATLAB is used for the development for this paper. It is chosen because it provides the Image Processing Toolbox and offers convenience to the development of the prototype due to its simple visualization of data and ease of development. The experiments are conducted on PC with the specifications as shown in Table 1.

Table 1. Specification of PC for experiments

Specification	Details
Processor	Intel Core 2 Duo P8400 @ 2.26 GHz, 3 MB L2 Cache
Motherboard	Acer Aspire 4935G, Intel PM45 + ICH9
Memory	Kingston 4 GB (2 × 2 GB) DDR2-SDRAM @ 800 MHz
System Drive	Western Digital 320 GB SATA 3Gb/s
Graphics	Nvidia GeForce 9300M GS, 512 MB
Operating System	Microsoft Windows XP Professional, 32-bit, Service Pack 3
Test environment	MATLAB R2011a

3.2 Raptor Flock Dataset

Videos of bird flocks are collected to be used for the experiment of this paper. Because raptors are bigger in size and are easier to be captured on video, the videos for the experiment are collected at Tanjung Tuan (Cape Rachado), Malacca, Malaysia during Raptor Watch Week 2011 on 12th to 13th March 2011 [12]. The original videos are taken using Panasonic Lumix FZ-100 with the resolution of 1080 × 960.

Six footages that were recorded are selected as the video dataset for the experiments that consists of flocks of Oriental Honey Buzzards (*Pernis ptilorhyncus*). All the videos are taken from a static position but with minor movements due to shake, wind that affects the movement of the camera. In order to provide a faster video processing speed, the original videos are resized to two smaller resolutions: 320 × 200 and 500 × 400. The details of the six videos are shown in Table 2.

Table 2. Conditions of the videos used in the experiment.

Video	Frames	Significant Cloud Outlines	Camera Movements	Raptor Distance to Camera
1	2640	Heavy portions	Mild jitters only	Far
2	2647	Heavy portions	Heavy movements and jitters	Far
3	573	None	None	Closer
4	3006	Small portions	Mild jitters only	Far
5	1471	None	Heavy movements and zoom out	Variable
6	1203	None	None	Near

4 Experimental Results and Analysis

Two experiments are conducted in this paper. The first experiment is to evaluate the accuracy of the tracking and counting algorithm. The second experiment is to evaluate the processing speed of the algorithm.

4.1 Experiment 1: Accuracy Evaluation

This experiment is to evaluate the accuracy of the proposed methodology. A raptor is considered to be successfully tracked if a bounding box is generated for it for at least 90% of its appearance in the video. The experimental results for the video resolution of 320 × 200 and 500 × 400 are shown in Table 3 and 4 respectively.

Table 3. Experimental results for video resolution of 320 × 200

Video	Raptor No.	Tracked	Hit Rate (%)	FAR (%)	Precision Rate (%)
1	21	15	71.4	69.26	21.96
2	19	16	84.2	56.85	36.33
3	12	9	75.0	5.23	71.08
4	18	11	61.1	13.46	52.89
5	36	25	74.2	10.24	62.33
6	23	17	73.9	2.44	72.11
Average Rate:			73.3	26.25	52.78

Table 4. Experimental results for video resolution of 500 × 400

Video	Raptor No.	Tracked	Hit Rate (%)	FAR (%)	Precision Rate (%)
1	21	18	85.7	72.13	23.80
2	19	18	94.7	58.45	39.36
3	12	11	91.6	5.29	86.81
4	18	15	83.3	15.43	70.48
5	36	31	86.1	12.37	75.46
6	23	20	86.9	2.89	84.44
Average Rate:			88.1	27.76	63.39

Based on the experimental results, it is shown that the video with the higher resolution provides a better hit rate, because when the resolution is too low, some of the higher flying raptors may be too small and are removed as noise during the process. It is shown that some small raptors that appears as small dots are not tracked in the video resolution of 320 × 200 but is successfully tracked in the video with a larger resolution of 500 × 400, as shown in Figure 4 where the red circles showed that the small raptors are only tracked in the video with resolution of 500 × 400 but not in 320 × 200.

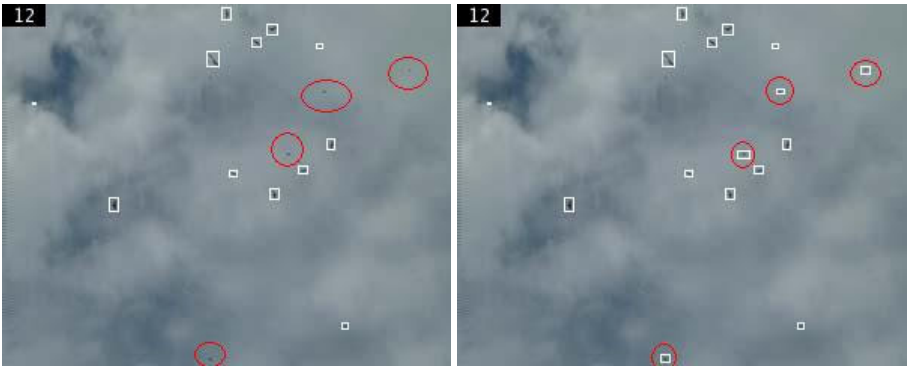


Fig. 4. Four raptors in red circles are not tracked in the video of resolution 300 × 200 (left) but are tracked in video of resolution 500 × 400 (right)

Fig. 5. An example of a false detection marked with a red circle on the top left, despite the absence of raptors

However, when it comes to the false acceptance rates (FAR), videos containing background with intense clouds with significant outline, coupled with slight movement of the camera could generate very high amount of false detections, which are exceptionally serious in the first two videos. This is because despite the morphological operations are performed on the video, the cloud's outlines are too solid than anticipated and the clouds are usually moving, causing the algorithm to treat it as part of a raptor as shown in Figure 5 where an area with outline of clouds is treated as a moving raptor.

4.2 Experiment 2: Speed Evaluation

This experiment is to evaluate the time performance of the proposed methodology under different video resolutions. Generally, the bigger the resolution, the longer it

should takes for the video to be processed. Table 5 shows the difference of time performance for the videos of the two different resolutions that are represented in frames per second (fps).

Table 5. Comparison of time performance for videos of different resolution (fps)

Video	Number of frames	320 × 200 (fps)	500 × 400 (fps)
1	2640	16.18	6.48
2	2647	16.90	6.70
3	573	15.85	4.08
4	3006	16.77	6.70
5	1471	16.69	5.48
6	1203	21.39	5.72
Average Rate:		16.98	6.17

Based on the experimental results, the video with resolution of 500 × 400 is significantly slower by 63.66% than the video with resolution of 320 × 200. However, both of them are not fast enough to be operated in real-time processing that requires at least a speed of at least 20-30 fps. However the video of 320 × 200 has an average speed of 16.98 and may be able to be further improved to achieve real-time calculations.

5 Conclusion

This paper is aimed to study the feasibility of automating the process of tracking multiple birds in flight within an area of viewing which was proven to be challenging for a human viewer to follow. The experimental results had shown that the hit rate could achieve 88.1% for video of 500 × 400 but in the mean time, very high FAR of up to 72.13% for the first video. But the third video is shown to be able to boost the precision rate to 86.81% and a high hit rate of 91.6%. Therefore, showing potential that if the FAR is successfully brought down, it would further boost the precision rate of the algorithm that is currently at an average of 63.39% that should have been further improved.

One of the major problems identified is the existence of moving clouds with significant boundaries that contrast with the blue skies strongly and therefore mistakenly accepted as moving raptors and causes an extremely high FAR in videos where such condition is heavy.

For the time performance, it is however too low for the video of 500 × 400 to process which only gives an average processing speed of 6.17 fps which was too slow to be implemented for real-time processes. The smaller video of 320 × 200 is performing faster at 16.98 fps but is still a gap away from the acceptable real-time processing speed requirements.

In the future, the filtering method should be better defined to filter off moving clouds, potentially with the use of simple verification on the shape of the raptor. The

silhouette could be extracted from the blobs and compared against potential silhouette of a flying raptor which might help to further reduce the FAR. It is also useful to analyze the flying path of the detected blob to check if it is an authentic bird, for example, if certain movement is hovering around a small area, it may had only been moving clouds in the background.

References

1. Lim, C.K., Earl of Cranbrook: Swiftlets of Borneo: Builders of Edible Nests. Natural History Publications (Borneo), Kota Kinabalu (2002)
2. Bildstein, K.L.: Migrating Raptors of the World: Their Ecology & Conservation. Comstock Publication Associates, Ithaca (2006)
3. Lin, W.H., Zheng, S.W.: A Field Guide to the Raptors of Taiwan. Yuan-Liou Publishing, Taiwan (2006)
4. DeCandido, R., Kasorndorkbua, C., Nualsri, C., Chinuparawat, C., Allen, D.: Raptor Migration in Thailand. In: Birding ASIA, vol. 10, pp. 16–22. Oriental Bird Club, UK (2008)
5. Yong, D.L.: An introduction to the Raptors South East Asia, Status, Identification, Biology and Conservation. Nature Society Singapore Bird Group (2010)
6. Gonzales, R.C., Woods, R.E.: Digital Image Processing, 3rd edn. Prentice Hall (2008)
7. Hilsmann, A., Eisert, P.: Optical Flow Based Tracking and Retexturing of Garments. In: Proc. 15th IEEE International Conference on Image Processing, ICIP, pp. 845–858 (2008)
8. Bloisi, D., Iocchi, L., Leone, G.R., Pigliacampo, R.: A Distributed Vision System for Boat Traffic Monitoringinthe Venice Grand Canal. In: Proc. of 2nd International Conference on Computer Vision Theory and Applications (VISAPP), pp. 549–556 (2007)
9. Zhang, X.Y., Wu, X.J., Zhou, X., Wang, X.G., Zhang, Y.Y.: Automatic Detection and Tracking of Maneuverable Birds in Videos. In: Proc. 2008 International Conference on Computational Intelligence and Security, pp. 185–188 (2008)
10. Baker, S., Matthews, I.: Lucas-Kanade 20 Years On: A Unifying Framework. International Journal of Computer Vision, LVI 3, 221–225 (2004)
11. Al Kanawathi, J., Mokri, S., Ibrahim, N., Hussain, A., Mustafa, M.: Motion Detection using Horn Schunck Algorithm and Implementation. In: Proc. International Conference Electrical Engineering and Informatics 2009 (ICEEI 2009), pp. 83–87 (2009)
12. Raptor Watch, `http://www.raptorwatch.org`

Author Index