Lecture Notes in Artificial Intelligence 7802

Subseries of Lecture Notes in Computer Science

Ali Selamat Ngoc Thanh Nguyen
Habibollah Haron (Eds.)

Intelligent Information and Database Systems

5th Asian Conference, ACIIDS 2013
Kuala Lumpur, Malaysia, March 18-20, 2013
Proceedings, Part I

Springer

Volume Editors

Ali Selamat
Universiti Teknologi Malaysia
Faculty of Computing, Department of Software Engineering
81310 UTM Skudai, Johor, Malaysia
E-mail: aselamat@utm.my

Ngoc Thanh Nguyen
Wrocław University of Technology
Institute of Informatics, Division of Knowledge Management Systems
Str. Wybrzeże Wyspiańskiego 27, 50-370 Wrocław, Poland
E-mail: ngoc-thanh.nguyen@pwr.edu.pl

Habibollah Haron
Universiti Teknologi Malaysia
Faculty of Computing, Department of Computer Science
81310 UTM Skudai, Johor, Malaysia
E-mail: habib@utm.my

ISSN 0302-9743 e-ISSN 1611-3349
ISBN 978-3-642-36545-4 e-ISBN 978-3-642-36546-1
DOI 10.1007/978-3-642-36546-1
Springer Heidelberg Dordrecht London New York

Library of Congress Control Number: 2013930842

CR Subject Classification (1998): I.2.1, I.2.3-4, I.2.6-8, I.2.11, H.3.3-5, H.2.8,
H.4.1-3, H.5.3, I.4.9, I.5.1-4, K.4.4

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 2013 was the fifth 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 2013 was to provide an internationally respected forum for scientific research in the technologies and applications of intelligent information and database systems. ACIIDS 2013 was co-organized by Universiti Teknologi Malaysia (Malaysia) and Wroclaw University of Technology Poland (Poland) in co-operation with Nguyen Tat Thanh University (Vietnam) and took place in Kuala Lumpur (Malaysia) during March 18–20, 2013. The first two events, ACIIDS 2009 and ACIIDS 2010, took place in Dong Hoi City and Hue City in Vietnam, respectively. The third event, ACIIDS 2011, took place in Daegu (Korea), while the fourth event, ACIIDS 2012, took place in Kaohsiung (Taiwan).

Submissions came from 20 countries from all over the world. Each paper was peer reviewed by at least two members of the International Program Committee and International Reviewer Board. Only 108 papers with the highest quality were selected for oral presentation and publication in the two-volumes proceedings of ACIIDS 2013.

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 new 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, Zaini Ujang (Universiti Teknologi Malaysia, Malaysia) and Tadeusz Więckowski (Rector of Wroclaw University of Technology, Poland) for their support.

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

We also would like to express our thanks to the Keynote Speakers (Hoang Pham, Naomie Salim, Mong-Fong Horng, Sigeru Omatu) for their interesting and informative talks of world-class standard.

We cordially thank our main sponsors, Universiti Teknologi Malaysia (Malaysia), Wroclaw University of Technology (Poland), and Nguyen Tat Thanh University (Vietnam). Our special thanks are due also to Springer for publishing the proceedings, and to the 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: Habibollah Haron (Organizing Chair) and the members of the Local Organizing Committee for their excellent work.

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

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

March 2013

Ngoc Thanh Nguyen
Ali Selamat

Conference Organization

Honorary Chairs

Zaini Ujang	Vice Chancellor of Universiti Teknologi Malaysia, Malaysia
Tadeusz Wieckowski	President of Wroclaw University of Technology, Poland

General Co-chairs

Ngoc Thanh Nguyen	Wroclaw University of Technology, Poland
Mohd Aizaini Maarof	Universiti Teknologi Malaysia, Malaysia

Program Chairs

Ali Selamat	Universiti Teknologi Malaysia, Malaysia
Shyi-Ming Chen	National Taiwan University of Science and Technology, Taiwan

Organizing Chair

Habibollah Haron	Universiti Teknologi Malaysia, Malaysia

Session Chairs

Andri Mirzal	Universiti Teknologi Malaysia, Malaysia
Bogdan Trawinski	Wroclaw University of Technology, Poland
Jason J. Jung	Yeungnam University, Republic of Korea

Organizing Committee

Roliana Ibrahim	Universiti Teknologi Malaysia, Malaysia
Nor Erne Nazira Bazin	Universiti Teknologi Malaysia, Malaysia
Dewi Nasien	Universiti Teknologi Malaysia, Malaysia
Mohamad Shukor Talib	Universiti Teknologi Malaysia, Malaysia

Steering Committee

Ngoc Thanh Nguyen	Chair, Wroclaw University of Technology, Poland
Longbing Cao	University of Technology Sydney, Australia
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
Toyoaki Nishida	Kyoto University, Japan
Leszek Rutkowski	Technical University of Czestochowa, Poland
Dickson Lukose	Knowledge Technology Cluster at MIMOS BHD, Malaysia

Keynote Speakers

Hoang Pham	Rutgers, The State University of New Jersey, USA
Naomie Salim	Universiti Teknologi Malaysia, Malaysia
Mong-Fong Horng	National Kaohsiung University of Applied Sciences, Taiwan
Sigeru Omatu	Osaka Institute of Technology, Japan

Special Sessions Organizers

1. *International Workshop on Engineering Knowledge and Semantic Systems (IWEKSS 2013)*

Jason J. Jung	Yeungnam University, Korea
Dariusz Krol	Wroclaw University of Technology, Poland

2. *Modelling and Optimization Techniques in Information Systems, Database Systems and Industrial Systems (MOT-ACIIDS 2013)*

Le Thi Hoai An	University of Lorraine, France
Pham Dinh Tao	INSA-Rouen, France

3. *Intelligent Supply Chains (ISC 2013)*

Arkadiusz Kawa	Poznan University of Economics, Poland
Paulina Golińska	Poznan University of Technology, Poland
Milena Ratajczak-Mrozek	Poznan University of Economics, Poland

4. *Intelligent Systems for Medical Applications (ISMA 2013) Information Systems and Industrial Engineering (MOT-ISIE)*

Uvais Qidwai	Qatar University, Qatar

5. *Innovations in Intelligent Computation and Applications (ICA 2013)*

Shyi-Ming Chen	National Taiwan University of Science and Technology, Taiwan

6. *Computational Biology and Bioinformatics (CBB 2013*)

Mohd Saberi Mohamad	Universiti Teknologi Malaysia, Malaysia

7. *Multiple Model Approach to Machine Learning (MMAML 2013)*

Tomasz Kajdanowicz	Wroclaw University of Technology, Poland
Tomasz Łuczak	Wroclaw University of Technology, Poland
Grzegorz Matoga	Wroclaw University of Technology, Poland

8. *Intelligent Recommender Systems (IRS 2013)*

Adrianna Kozierkiewicz-Hetmańska	Wroclaw University of Technology
Ngoc Thanh Nguyen	Wroclaw University of Technology

9. *Applied Data Mining for Semantic Web (ADMSW 2013)*

Trong Hai Duong	Quang Binh University, Vietnam
Bay Vo	Information Technology College, Vietnam

International Program Committee

Abdul Rahim Ahmad	Universiti Tenaga Nasional, Malaysia
Abdul Samad Ismail	Universiti Teknologi Malaysia, Malaysia
Abdul Samad Shibghatullah	Universiti Teknikal Malaysia, Malaysia
Adrianna Kozierkiewicz-Hetmańska	Wrocław University of Technology, Poland
Alex Sim	Universiti Teknologi Malaysia, Malaysia
Alvin Yeo	Universiti Malaysia Sarawak, Malaysia
Amir Shafie	International Islamic University Malaysia, Malaysia
Annabel Latham	The Manchester Metropolitan University, UK

Antoni Wibowo	Universiti Teknologi Malaysia, Malaysia
Arkadiusz Kawa	Poznan University of Economics, Poland
Aryati Bakri	Universiti Teknologi Malaysia, Malaysia
Azlan Mohd Zain	Universiti Teknologi Malaysia, Malaysia
Azurah Abu Samah	Universiti Teknologi Malaysia, Malaysia
Bay Vo	Information Technology College, Vietnam
Behnam Rouzbehani	Islamic Azad University, Central Tehran Branch, Iran
Bing-Han Tsai	National Taiwan University of Science and Technology, Malaysia
Bjoern Schuller	Technische Universität München, Germany
Bogdan Trawinski	Wroclaw University of Technology, Poland
Boguslaw Cyganek	AGH University of Science and Technology, Poland
Cheng-Yi Wang	National Taiwan University of Science and Technology, Taiwan
Dariusz Barbucha	Gdynia Maritime University, Poland
Dariusz Frejlichowski	West Pomeranian University of Technology, Poland
Dariusz Krol	Bournemouth University, UK
Dongjin Choi	Chosun University, Korea
Dragan Simic	University of Novi Sad, Serbia
El-Houssaine Aghezzaf	Ghent University, Belgium
Elżbieta Kukla	Wrocław University of Technology, Poland
Eric Pardede	La Trobe University, Australia
Faisal Zaman	Kyushu Institute of Technology, Poland
Fan Wang	Microsoft, USA
Geetam Tomar	Malwa Institute of Technology and Management, Gwalior, India
Gia-An Hong	National Taiwan University of Science and Technology, Taiwan
Gordan Jezic	University of Zagreb, Coatia
Hae Young Lee	ETRI, Korea
Halina Kwasnicka	Wroclaw University of Technology, Poland
Hoai An Le Thi	University of Lorraine, France
Huey-Ming Lee	Chinese Culture University, Taiwan
Huynh Binh	University of Science and Technology, Vietnam
Hyon Hee Kim	Dongduk Women's University, Korea
Imran Ghani	Universiti Teknologi Malaysia, Malaysia
Ireneusz Czarnowski	Gdynia Maritime University, Poland
Iskandar Ishak	Universiti Putra Malaysia, Malaysia
Jafar Razmara	Universiti Teknologi Malaysia, Malaysia

Jaroslaw Jankowski	West Pomeranian University of Technology in Szczecin, Poland
Jason Jung	Yeungnam University, Korea
Jerome Euzenat	INRIA, France
Jesús Alcalá-Fdez	University of Granada, Spain
Jose Norbeto Mazon	University of Alicante, Spain
Kamal Zamli	Universiti Malaysia Pahang, Malaysia
Kang-Hyun Jo	University of Ulsan, Korea
Katarzyna Grzybowska	Poznan University of Technology, Poland
Kazuhiro Kuwabara	Ritsumeikan University, Japan
Khairuddin Omar	Universiti Kebangsaaan Malaysia, Malaysia
Lian En Chai	Universiti Teknologi Malaysia, Malaysia
Manh Nguyen Duc	ENSTA Bretagne, France
Marcin Hajdul	Institute of Logistics and Warehousing, Poland
Maria Bielikova	University of Technology in Bratislava, Slovakia
Md. Nazrul Islam	Universiti Teknologi Malaysia, Malaysia
Michał Woźniak	Wroclaw University of Technology, Poland
Milena Ratajczak-Mrozek	Poznan University of Economics, Poland
Minh Le Hoai	University of Lorraine, France
Mohd Helmy Abd Wahab	Universiti Tun Hussein Onn Malaysia, Malaysia
Mohd Murtadha Mohamad	Universiti Teknologi Malaysia, Malaysia
Mohd Ramzi Mohd Hussain	International Islamic University Malaysia, Malaysia
Mohd Saberi Mohamad	Universiti Teknologi Malaysia, Malaysia
Moon II Chul	Korean Institute Science and Technology, Korea
Muhammad Khan	King Saud University, Saudi Arabia
Muhamad Razib Othman	Universiti Teknologi Malaysia, Malaysia
Mohd Salihin Ngadiman	Universiti Teknologi Malaysia, Malaysia
Muhammad Shuaib Karim	Quaid-i-Azam University, Malaysia
Muhammad Suzuri Hitam	Universiti Malaysia Terengganu, Malaysia
Nghi Do Thanh	Telecom-Bretagne, France
Ngoc Thanh Nguyen	Wroclaw University of Technology, Poland
Niels Pinkwart	Clausthal University of Technology, Germany
Nojeong Heo	Dongyang University, Korea
Noorfa Haszlinna Mustaffa	Universiti Teknologi Malaysia, Malaysia
Nor Azizah Ali	Universiti Teknologi Malaysia, Malaysia
Nor Haizan Mohamed Radzi	Universiti Teknologi Malaysia, Malaysia
Nor Hawaniah Zakaria	Universiti Teknologi Malaysia, Malaysia

Norazah Yusof	Universiti Teknologi Malaysia, Malaysia
Norazman Ismail	Universiti Teknologi Malaysia, Malaysia
Olgierd Unold	Wroclaw University of Technology, Poland
Ondrej Krejcar	University of Hradec Kralove, Czech Republic
Trong Hai Duong	Inha University, Korea
Bernadetta Mianowska	Wroclaw University of Technology, Poland
Michal Sajkowski	Poznan University of Technology, Poland
Robert Susmaga	Poznan University of Technology, Poland

Table of Contents – Part I

Innovations in Intelligent Computation and Applications -1

Innovations in Intelligent Computation and Applications -2

Intelligent Database Systems -1

Intelligent Database Systems -2

Intelligent Information Systems -1

Intelligent Information Systems -2

Intelligent Information Systems -3

Table of Contents – Part II

Tools and Applications

Intelligent Recommender Systems

Multiple Model Approach to Machine Learning

Engineering Knowledge and Semantic Systems

Computational Biology and Bioinformatics

Computational Intelligence

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

Intelligent Supply Chains

Applied Data Mining for Semantic Web

Semantic Web and Ontology

Integration of Information systems

Conceptual Modeling in Advanced Database Systems

Intelligent Moving Objects Detection via Adaptive Frame Differencing Method

Chun-Ming Tsai[1,*] and Zong-Mu Yeh[2]

[1] Department of Computer Science, Taipei Municipal University of Education,
No. 1, Ai-Kuo W. Road, Taipei 100, Taiwan
cmtsai2009@gmail.com

[2] Department of Mechatronic Technology, National Taiwan Normal University,
Taipei 106, Taiwan
zongmu@ntnu.edu.tw

Abstract. The detection of moving objects is a critical first step in video surveillance, but conventional moving objects detection methods are not efficient or effective for certain types of moving objects: slow and fast. This paper presents an intelligent method to detect slow- and fast-moving objects simultaneously. It includes adaptive frame differencing, automatic thresholding, and moving objects localization. The adaptive frame differencing uses different inter-frames for frame differencing, the number depending on variations in the differencing image. The thresholding method uses a modified triangular algorithm to determine the threshold value and reduces most small noises. The moving objects localization uses six cascaded rules and bounding-boxes-based morphological operations to merge broken objects and remove noise objects. The fps value (maximum 72) depends on the speed of the objects. The number of inter-frames is inversely proportional to the speed. The results demonstrate that our method is more efficient than traditional frame differencing and background subtraction methods.

Keywords: Adaptive frame differencing, Moving objects Detection, Video surveillance, Bounding-boxes-based morphological operations, background subtraction.

1 Introduction

Moving objects detection (MOD) is the most important step in a video surveillance system [1]. MOD methods fall into four main types [2]: background subtraction (BS) [3], optical flow (OF) [4], frame differencing (FD) [5] and hybrid methods [6]. BS methods detect moving objects as the difference between the current frame and an image of the background model. However, there are many problems encountered in the BS methods: (1) automatic building of the background model image; (2) updating the background model; and (3) deciding the update rate. Researchers have proposed

[*] This paper is supported by the National Science Council, R.O.C., under Grants NSC 100-2221-E-133-004- and NSC 101-2221-E-133-004-.

A. Selamat et al. (Eds.): ACIIDS 2013, Part I, LNAI 7802, pp. 1–11, 2013.

many methods to solve these problems [7-9]. OF methods use the motion of the moving objects to produce intensity changes in the magnitude, which become important cues for locating the object in time and space. However, their relationship is not unique due to the presence of noise or other external causes like illumination drifts. And these algorithms are not computationally efficient [8]. FD methods based on the difference between two or three consecutive frames provide the simplest method for detecting moving objects and can adapt to dynamic environments. However, FD methods cannot detect the entire shape of the moving object [6] and they are very sensitive to the threshold value used to convert the differencing image to a binarization image. Hybrid methods [6] combine the other methods, but they cannot solve the problems of BS and OF methods.

In this paper, we will propose an efficient and effective MOD method to solve the problems encountered in traditional FD methods. These solutions include (1) using adaptive frame differencing; (2) obtaining a robust difference threshold value; (3) extraction of the entire moving object; and (4) developing a method for processing moving objects that move slowly and quickly.

2 Adaptive Frame Differencing (AFD)

Take a video containing 1400 frames as an example. The two-frame differencing (2FD) method is used to obtain the differencing image. We compute the variance for each differencing image (Figure 1). The horizontal axis represents the index of the 2FD image; the vertical is the variance for each differencing image. From Fig. 1, it is observed that several of the variances are zero. For example, ten variances between 0 and 200 differencing images are zero; some between 1200 and 1400 are very significantly large; and some between 0 and 71 are moderately large. By contrast, many variances are insignificant, e.g., between 72 and 647 (below 10.0). Based on these observations, the representation for these various variances and the method of processing these video frames are described.

First, very small variances of the differencing images it implies that the two successive frames are similar. Thus, for this case, the current frame needs not to be processed; we skip to the next frame. Secondly, if the variances of the differencing images are large or moderate, the change between the two successive frames is significant, the speed of the moving objects is high, or the distance between the moving objects and the video camera is small. In these cases, the changes can be easily detected by the conventional 2FD or 3FD method. However, if the variances are insignificant, it implies that the speed of the moving objects is slow. Finally, if the distance between the moving objects and the video camera is increased, the variances of the differencing images trends to decrease. Under these conditions, the following conventional procedures (removal of noises, binary morphological operations, and connected component labeling) cannot detect the moving objects properly. Moreover, both the time and space complexities of 3FD are thrice and twice than that of 2FD, respectively. Thus, 2FD is faster than 3FD, but its detection accuracy is less.

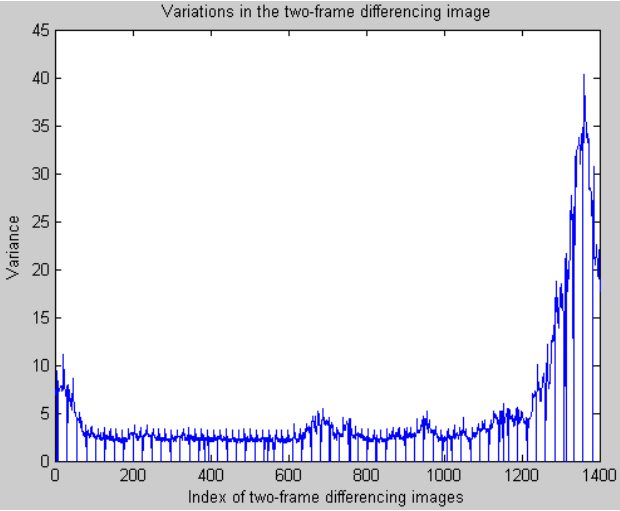


Fig. 1. Variations in the image formed by two-frame differencing

Hence, in this paper, a general FD algorithm is proposed to detect the moving objects intelligently, based on the variation in the differencing images. If the moving object is fast-moving (or if the distance between the moving objects and the video camera is small,) the variation in the difference image is large, while if the object is slow-moving (or if it is far from the video camera) is far, the variation in the difference images is small. The main formula of the proposed adaptive frame differencing (AFD) technique is

$$D_i(x, y) = \left| F_i(x, y) - F_{i-n}(x, y) \right| > T_i \ . \qquad (1)$$

Here, D_i is the difference image, i is the current frame image of the input video, n is the interval between the next chosen frame and the current frame, and T_i is a variable threshold adaptive to i, described in Section 3. The algorithm of the proposed AFD is described as follows.

(1) Set n, the initial the interval frame number as 1.
(2) Similarity: If the difference mean between the ith frame image and the (i-n)th is smaller than a pre-learning parameter, these two images are deemed to be similar.
(3) If step (2) is true, go to the next frame, and repeat step (2).
(4) Use Eq. (1) to obtain difference image, D_i.
(5) Sufficiency: If the variation in D_i is larger than a pre-learning parameter ($T_{i,}$), the variation is sufficient.
(6) If step (5) is false, go to the next frame and increase the interval frame number (n); and then repeat step (2).
(7) If step (5) is true, continue automatic thresholding method.

From the above AFD algorithm, it is apparent that step 2 is used to detect whether the two frames are similar. If they are, the successive frame can be skipped and the next frame processed, reducing processing time. Step 5 is used to detect whether the variation in the differencing image is sufficient. A significant variation implies that the moving object is fast-moving or its distance from the video camera is small. In these cases, the number of the interval frame is 1. If the variation is insignificant, it reflects that the motion object is slow-moving or it is distant from the video camera. In these cases, the number of the interval frame is increased, and we repeat the steps. The smaller is the variation of the differencing image, the greater the number of frames in the interval. Hence, number of the interval frame (n) is adaptive to the video content, and the proposed method is named "AFD" algorithm.

3 Automatic Thresholding (AT)

To segment the moving objects, a differencing image is converted into a binary image, and connected component labeling (CCL) is used to segment the moving objects. However, if the threshold value is too small, excessive noise is produced in the binary image. Conventional MOD methods usually employ binary morphological (BM) operations to remove the noise. However, with an improper threshold value, the connected components cannot be extracted properly by removing the noise, and if the threshold value is too large, many broken objects are produced. Thus, selecting proper threshold value is very important.

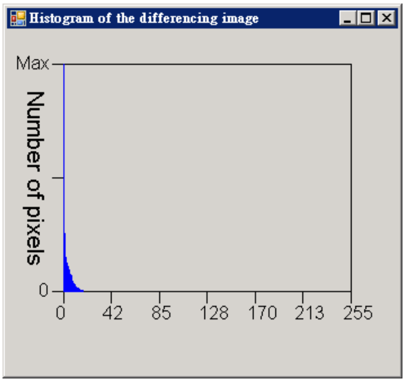


Fig. 2. Histogram of the differencing image

In the histogram of the differencing image (Fig. 2), the horizontal axis is the gray level of the differencing image, and the vertical axis is the number of pixels with each gray level in the differencing image. If two frames are similar, the variation in the differencing image is located at a low gray level; if they are dissimilar, the variation is located at a high gray level. Moreover, one large peak at the low gray level usually indicates that most of the two images are stationary. The greater the change, higher is the gray level in the differencing image, and differencing images usually do not have two normal distributions. Thus, it is improper to classify this differencing image into two classes by the Otsu thresholding method [10]. Instead, herein, the modified triangle thresholding (MTT) method [11], described below, is used to threshold a differencing image into a binary image.

In the conventional FD method, binary morphological (BM) operations are used to remove the noise but these operations must repeatedly process the entire image. Herein, the automatic thresholding (AT) method is employed with a larger threshold value to produce binary result with little noise. Next, CCL is used directly to detect the moving objects. The BM operations are not used before implementing the CCL. Notably, this procedure speeds up the entire process. The proposed AT algorithm is described as follows.

(1) Obtain the histogram of the differencing image.
(2) Use a Gaussian smoothing filter [11] to smooth the original histogram to remove noise.
(3) Use the MTT method [11] to obtain the threshold value D_{thr}.
(4) Obtain the automatic threshold value T_i by the equation

$$T_i = D_{thr} + k \times \sigma \; . \tag{2}$$

Here, k is a noise removal constant and is used to adjust the threshold value to obtain a binary image with less noise (herein, k is set as 2.0 by using pre-learning method). σ is the standard deviation of the differencing image and depends on the surveillance environment.

(5) Use T_i to convert the differencing image to the binary image.

From the above AT algorithm, when the variation between two frames is large, the variance of the differencing image is also large. Thus, the automatic threshold value is adjusted to a higher limit to remove more noise and reserve the moving objects. If the difference between two frames is small, the variance of the differencing image is also small, and the automatic threshold value is adjusted lower to convert the differencing image into a binary image.

4 Cascaded Moving Objects Localization

After the AT and CCL steps, many moving objects have been extracted. These moving objects are represented by the bounding-boxes (BBs), which may be humans, cars, or noise. Herein, a cascaded localization method is proposed to remove small and spread noise BBs, to merge concentrated broken BBs, and to locate the moving objects. The proposed localization method includes six cascaded rules, described as follows.

(1) Removing light shadow BBs: The characteristic feature of light shadows is that a large variation in their luminance mean u between the previous image and the current frame image, so we use u to remove noise produced by these light shadows, as follows: If the difference in luminance mean between the previous and current frames in a BB is greater than a predefined value (T_u) (set from learning as 50), BB-based erosion [12] is used to remove this BB.
(2) Merging concentrated BBs to form a complete object: If some BBs are centralized, the distance feature is used to merge the concentrated BBs into a

large BB, by a rule defined as follows: If the distance between two BBs is smaller than a predefined value (T_{D1}), BB-based closing [12] is used to merge them. T_{D1} is set to 0.0125 * *min* (*W*, *H*) obtained from learning. *W* and *H* are the width and height of the video frame, respectively.

(3) Removing scattered and small-area BBs (random noise) that remain after the previous two steps by applying area and distance features, by a rule as: If the area of BB is smaller than a predefined value (T_{A1}) and the distance between this BB and another BB is greater than a predefined value (T_{D2}), BB-based opening [12] is used to remove it. T_{A1} and T_{D2} are set to 0.00004 * *W** *H* and 0.1 * *min* (*W*, *H*), respectively, obtained from learning.

(4) Merging small and concentrated BBs that are produced by a moving object whose motion is insignificant. In particular, a CCD camera can monitor many people, some of whom make insignificant motions, while others move significantly. The insignificant motion produces broken BBs after the threshold result. To merge such broken BBs, we use area and distance features by a rule is defined as follows: If the area of a BB is smaller than a predefined value (T_{A2}) and the distance between this BB and another BB is smaller than a predefined value (T_{D3}), a BB-based closing [12] process is used to merge them. T_{A2} and T_{D3} are set to 0.00125 * *W** *H* and 0.1 * *min* (*W*, *H*), respectively, obtained from learning.

(5) Removing erroneously merged BBs that contain small motion pixels produced by the merging rule, using the motion pixel ratio by a rule defined as follows: If a BB's motion pixel ratio is smaller than a predefined value (T_{MPR}), BB-based erosion [12] is used to remove it. T_{MPR} is set to 0.000027, obtained from learning.

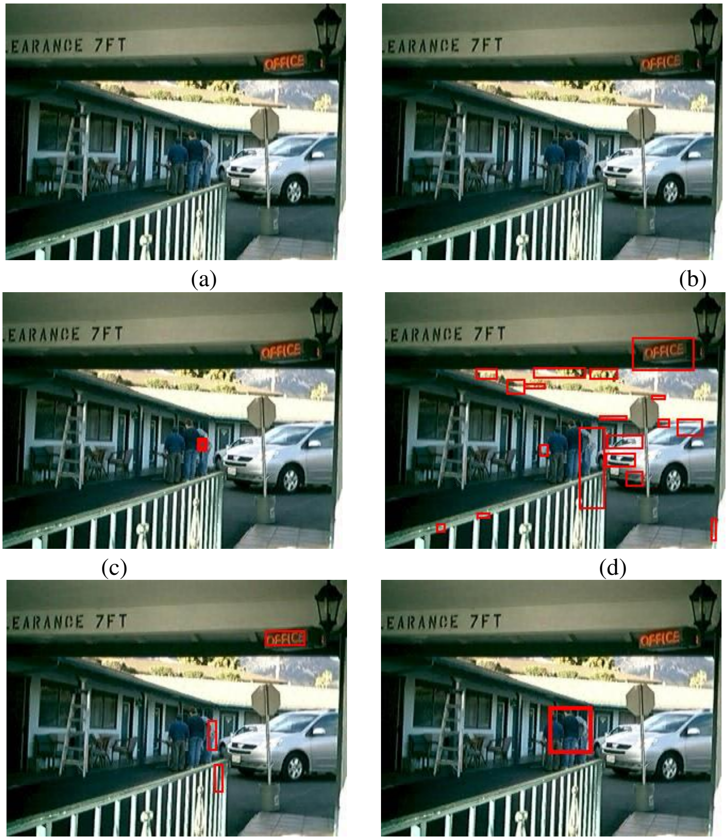
(6) Merging closed small BBs to merge the images with non-human shapes and closed small BBs. Applying the five abovementioned rules produces some BBs with non-human shape. In particular, when a human moves slowly, the movements of limbs are significant. However, the movements performed by the rest of his body are insignificant and a broken image of the body is produced. The aspect ratio and distance features are used to merge the broken body image, by a rule defined as follows: If the distance between two BBs is smaller than a predefined value (T_{D4}) and their aspect ratios do not conform to the ratio of the human shape, BB-based closing [12] is used to merge them. T_{D4} is set to 0.03 * *H*, obtained from learning.

5 Experimental Results

MOD results and performance analysis of the execution times are compared for 2FD, 3FD, and AFD methods. Further, two BS methods: *Mixture of Gaussians* (MoG) [13] and MoG with adaptive number of Gaussians (AMoG) [9] are used to compare the results.

For the video clips of slow-moving objects – three persons are in a conversation (Fig. 3)--, the detection results of AFD, 2FD, and 3FD methods are shown in Table 1.

This video clip has 52 identical frames, the mean and variance of current and previous frames are equal, and the motions are insignificant. In this case, 2FD and 3FD methods are time-consuming and even so cannot detect the moving objects. Their false negative rates are high. The much lower false negative rate of the proposed AFD method is caused by an object that is almost stationary. From Table 1, it is evident that the proposed AFD method is superior to conventional FD methods.



(a) (b)

(c) (d)

(e) (f)

Fig. 3. Example of slow-moving objects detection. (a) #526; (b) #575; (c) detected by 2FD; (d) detected by MoG; (e) detected by AMoG; (f) detected by AFD

Table 1. Comparison of detection results obtained by using AFD, 2FD and 3FD for slow-moving video clips (320 × 240)

Methods	Total Objects	True Positive Rate	False Negative Rate	False Positive Rate
AFD	3862	**96**%	4%	1%
2FD	3862	73%	27%	3.4%
3FD	3862	65%	35%	2%

The performance time analyses of AFD, 2FD, 3FD, and AMoG are shown in Table 2. This comparison used a 1334-frame video clip, in which the moving objects' sizes are from 26×60 to 188×239 pixels, the maximum size being produced by the shortest distance between the moving object and the camera. To process this video clip, the execution times for AFD, 2FD, 3FD, AMoG are 18.457, 31.162, 40.874, and 115.149 seconds, respectively. The average frames per second (FPS) for AFD, 2FD, and 3FD are 72.27, 42.81, 32.64, and 11.58, respectively. The AFD method skips similar frames, and when the variance of the differencing image is small, the inter-frame is increased. Thus, the AFD method is the fastest method, and is be superior to the traditional FD and AMoG methods (Table 2).

Table 2. Comparison of time performance obtained by using AFD, 2FD and 3FD for slow-moving video clips (320 × 240)

Methods	Execution Times for 1334 frames video clip	FPS
AFD	18.457 (s)	**72.27**
2FD	31.162 (s)	42.81
3FD	40.874 (s)	32.64
AMoG	115.1489 (s)	11.58

For the objective evaluation, the first data set of the 2006 IPPR contest (http://140.109.20.238/) was adopted. This data set involves indoor surveillance and 456 moving objects whose sizes range from 4×8 to 75×147 pixels. When the distance between the moving object and the CCD camera is large, the object size is small, and conversely when the distance is short, the size of the object is large. In the first data set, the objects move at different speeds, some walk, some run, and others walk and remain stationary, alternately. The data set provides manually extracted ground truth data. Table 3 shows the detection results obtained by using AFD, 2FD, and 3FD, respectively, and shows that the AFD method is superior to the 2FD method at a low false positive rate. Further, the AFD method is superior to 3FD method at low false positive rates and high true positive rates. Table 4 shows the execution times obtained by using AFD, 2FD, 3FD, and AMoG for this data set, showing the execution times for AFD and 2FD methods to be similar. Both these methods process the salient motion object faster than 3FD and AMoG methods.

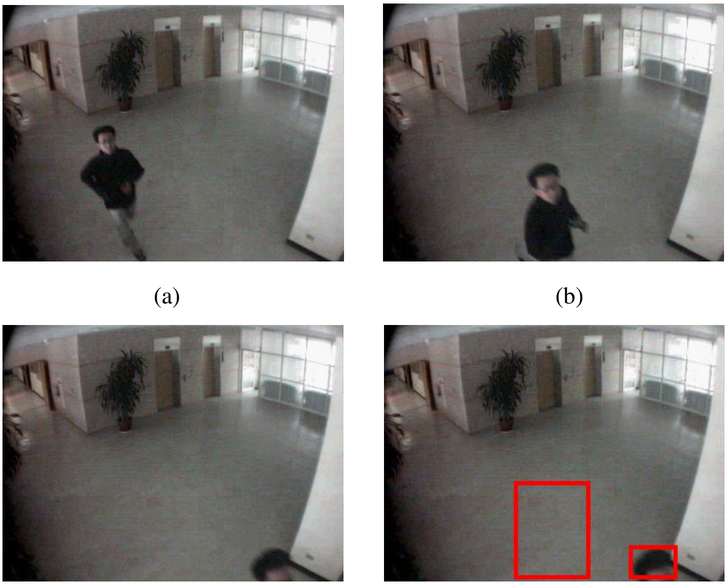
Table 3. Comparison of detection results obtained by using AFD, 2FD and 3FD for 2006IPPR contest's first data set (320 × 240)

Methods	Total Objects	True Positive Rate	False Negative Rate	False Positive Rate
AFD	456	**97.58**%	2.42%	0.438%
2FD	456	97.15%	2.85%	4.605%
3FD	456	92.98%	7.02%	5.92%

Table 4. Comparison of time performance obtained by using AFD, 2FD and 3FD for 2006IPPR contest's first data set (320 × 240)

Methods	Execution Times for 1334 frames video clip	FPS
AFD	6.592 (s)	**69.17**
2FD	7.247 (s)	62.92
3FD	9.257 (s)	49.26
AMoG	29.048(s)	10.33

In the video clip showing slow-moving objects, their motions are insignificant. In an example (Fig. 3), two men and a one woman are involved in a discussion. Their actions are not obvious. 2FD, 3FD, MoG, and AMoG methods cannot detect these slow-moving objects, but the proposed AFD method detects these slow-moving objects by using 49 inter-frames. Figures 3(a)–(b) indicate the original frames (#526 and #575). The detected results on frame 575 by 2FD, MoG, AMoG, and AFD are shown in Figs. 3(c)-(f), respectively.



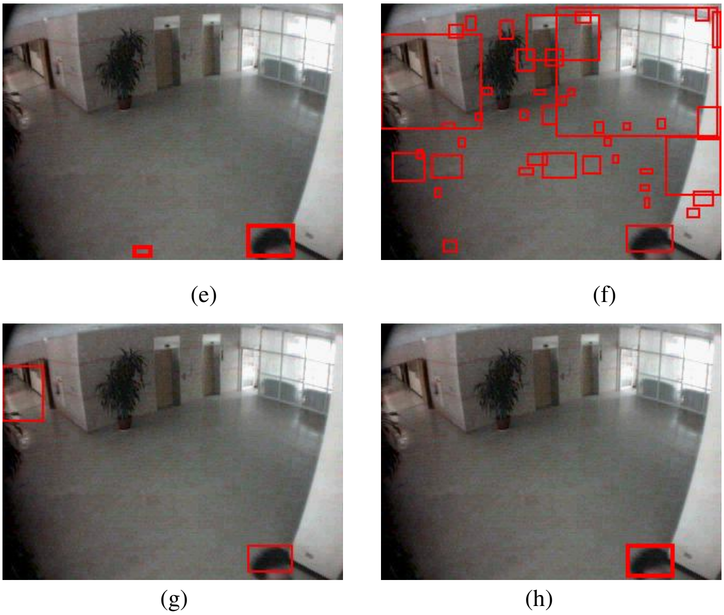
(a)

(b)

(c)

(d)

Fig. 4. Example of fast-moving objects detection. (a) #147; (b) #148; (c) #149 (d) detected by 2FD; (e) detected by 3FD; (f) detected by MoG; (g) detected by AMoG; (h) detected by AFD.



(e) (f)

(g) (h)

Fig. 4. (*Continued*)

The motions of fast-moving objects are salient. In an example (Fig. 4), a man is running. The original 147th, 148th, and 149th frame images are shown in Figs. 4(a)–(c), respectively. The detected results on frame 149 by 2FD, 3FD, MoG, AMoG, and AFD are shown in Figs. 4(d)-(h), respectively. The proposed AFD method can detect the fast-moving object, but the others produced "ghost" and noise objects.

More detailed results of the video clips can be viewed at http://cmtsai.tmue.edu.tw/~cmtsai/MOE/ACIIDS2013.htm

6 Conclusions

This study proposes an intelligent, efficient, and effective MOD method that includes an AFD, AT, cascaded rules, and BBM operations. Experimentally, slow-moving objects in a video clip can be detected by adaptively adjusting the number of the inter-frames. Furthermore, the proposed AFD method can skip similar frames to speed process time, and can determine the threshold value automatically. The experimental results show that the proposed method is computationally more efficient and effective than the traditional FD and BS methods.

References

1. Gao, T., Jia, Y.J., Wang, P., Wang, C.S., Zhao, J.T.: Feature particles tracking for traffic moving object. Adv. Sci. Lett. 5, 802–805 (2012)
2. Radke, R., Andra, S., Al-Kofahi, O., Roysam, B.: Image change detection algorithms: a systematic survey. IEEE Trans. on Image Processing 14(3), 294–307 (2005)

3. Piccardi, M.: ‘Background Subtraction Techniques: A Review. In: IEEE International Conference on SMC, vol. 4, pp. 3099–3104 (2004)
4. Velastin, S.A., Boghossian, B.A., Lo, B.P.L., Sun, J., Vicencio-Sliva, M.A.: PRISMATICA: toward ambient intelligence in public transport environments. IEEE Transactions on SMC - Part A 35(1), 164–182 (2005)
5. Kim, C., Hwang, J.: Fast and automatic video object segmentation and tracking for content-based applications. IEEE Trans. on CSVT 12(2), 122–129 (2002)
6. Tian, Y.L., Hampapur, A.: Robust salient motion detection with complex background for real-time video surveillance. In: Proc. of IEEE Computer Society Workshop on Motion and Video Computing, vol. 2, pp. 30–35 (2005)
7. Heikkila, M., Pietikainen, M.: A texture-based method for modeling the background and detecting moving objects. IEEE Transactions on PAMI 28(4), 657–662 (2006)
8. Wei, Z.Q., Ji, X.P., Wang, P.: Real-time moving object detection for video monitoring systems. Journal of System Engineering and Electronics 17(4), 731–736 (2006)
9. Zivkovic, Z., van der Heijden, F.: Efficient adaptive density estimation per image pixel for the task of background subtraction. Pattern Recognition Letters 27(7), 773–780 (2006)
10. Otsu, N.: A thresholding selection method from gray-scale histogram. IEEE Trans. SMC 9, 62–66 (1979)
11. Tsai, C.M., Lee, H.J.: Binarization of color document images via luminance and saturation color features. IEEE Transactions on Image Processing 11(4), 434–451 (2002)
12. Tsai, C.M.: Intelligent Post-processing via Bounding-Box-Based Morphological Operations for Moving Objects Detection. In: Jiang, H., Ding, W., Ali, M., Wu, X. (eds.) IEA/AIE 2012. LNCS, vol. 7345, pp. 647–657. Springer, Heidelberg (2012)
13. Stauffer, C., Grimson, W.E.L.: Adaptive background mixture models for real-time tracking. In: IEEE Computer Society Conference on CVPR, vol. 2, pp. 246–252 (1999)

Global Artificial Bee Colony Algorithm for Boolean Function Classification

Habib Shah, Rozaida Ghazali, and Nazri Mohd Nawi

Faculty of Computer Science and Information Technology
Universiti Tun Hussein Onn Malaysia (UTHM)
Parit Raja, 86400 Batu Pahat. Johor, Malaysia
habibshah.uthm@gmail.com, {rozaida,nazri}@uthm.edu.my

Abstract. This paper proposed Global Artificial Bee Colony algorithm for training Neural Network (NN), which is a globalised form of standard Artificial Bee Colony algorithm. NN trained with the standard backpropagation (BP) algorithm normally utilizes computationally intensive training algorithms. One of the crucial problems with the BP algorithm is that it can sometimes yield the networks with suboptimal weights because of the presence of many local optima in the solution space. To overcome, GABC algorithm used in this work to train MLP learning for classification problem, the performance of GABC is benchmarked against MLP training with the typical BP, ABC and Particle swarm optimization for boolean function classification. The experimental result shows that MLP-GABC performs better than that standard BP, ABC and PSO for the classification task.

Keywords: Artificial Bee Colony algorithm, Back propagation, Global Artificial Bee Colony algorithm.

1 Introduction

Artificial Neural Network (ANN) has the most novel and powerful tools suitable for solving combinatorial problems such as prediction, clustering and classification task [1-3]. It is being applied to different optimization and mathematical problems such as object, image recognition, signal processing and numeric function [4].

There are several techniques used in favourable performance for training NN such as evolutionary algorithms (EA), genetic algorithm (GA), PSO, differential evolution, ant colony optimization (ACO), BP and improved BP algorithm [5-10]. These techniques are used for initialization of best weights, parameters, activation function, and selection of a proper network structure.

The main task of BP algorithm is to update the network weights for minimizing output error [11]. It has a high success rate in solving many complex problems, but it still has some drawbacks, especially when setting parameter values like initial values of connection weights, value for learning rate, and momentum. If the network topology is not carefully selected, the NNs algorithm can get trapped in local minima, or it might lead to slow convergence or even network failure. In order to overcome the disadvantages of standard BP, much global optimization technique used.

A. Selamat et al. (Eds.): ACIIDS 2013, Part I, LNAI 7802, pp. 12–20, 2013.

ABC, Hybrid Ant Bee Colony (HABC), an Improved ABC and the Global Hybrid Ant Bee Colony algorithms, can provide the best possible solutions for mathematical problems [12-17]. A common feature of these algorithms is that the population consisting of feasible solutions to the difficulty is customized by applying some agents on the solutions depending on the information of their robustness. Therefore, the population is encouraged towards improved solution areas within the solution space. GABC algorithm is a population-based proposed here for training MLP. In this study, GABC algorithm is used successfully to train MLP for Boolean function classification.

2 Boolean Function Classification

Classification of data concern with the use of computers in order to create a structure that learns how automatically chooses to which of a predefined set of classes, a given object belongs. Boolean function classification is the most important issue until today how to decide the 0 and 1 or on or off separate classes by different techniques. These are Boolean functions such as XOR, 3-Bit Parity and Encoder Decoder problems. These are non-linear benchmark classification tasks consisting of 2^N patterns with N inputs and one output. Each input or output is either a 0 or a 1.The three tasks for classification, which are benchmarked problems used in training MLP using GHAB algorithm [1].

Problem 1: The XOR benchmark is a difficult classification problem mapping two binary inputs to a single binary output.

Problem 2: Benchmark 3-Bit Parity Problem: The problem is taking the modulus 2 of summation of three inputs. In other words, if the number of binary inputs is odd, the output is 1, otherwise it is 0.

Problem 3: 4-Bit Encoder/Decoder is quite close to real world pattern classification task, where small changes in the input pattern cause small changes in the output pattern [18-19].

3 Artificial Neural Networks (ANN)

MLP, which is also known as a feed forward neural network was introduced for the non-linear XOR, and was then successfully applied to different combinatorial problems [20]. It is mostly used for information processing and pattern recognition. MLP

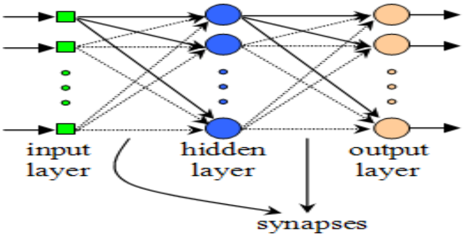


Fig. 1. Multilayer perceptron network

is highly used and tested with different problems such as in time series prediction and function approximation. Figure 1 shows the architecture of MLP with two hidden layers, one output layer, and one input layer.

$$y_i = f_i\left(\sum_{j=1}^{n} w_{ij} x_j + b_i\right) \tag{1}$$

where y_i is the output of the node, x_i is the j_{th} input to the node, w_{ij} is the connection weight between the input node and an output node, theta is the threshold (or bias) of the node, and f_i is the node transfer function. Usually, the node transfer function is a non-linear function such as a sigmoid function, a Gaussian function, and etc. The network error function E will be minimized as

$$E(w(t)) = \frac{1}{n}\sum_{j=1}^{n}\sum_{k=1}^{n}(d_k - o_k)^2 \tag{2}$$

where *E(w(t))* is the error at the tth iteration; *w(t)* is the weights in the connections at the tth iteration; d_k is the desired output node; O_k is the actual value of the k_{th} output node; k is the number of output nodes; and n is the number of patterns. Meanwhile, t is the optimization target to minimize the objective function by optimizing the network weights *w(t).*

BP is currently the most widely and well known used algorithm, for training MLP [11, 21]. The BP is a gradient descent method in which gradient of the error is calculated with respect to the weight's values for a given input by propagating the error backwards from the output to hidden layer and further to input layer continuously. This step by step mathematical procedure adjusts the weights according to the error function.

Although the BP algorithm is a powerful technique applied to MLP training. However, as the problem complexity increases, the performance of BP falls off rapidly because gradient search techniques tend to get trapped at local minima. When the nearly global minima are well hidden among the local minima, BP can end up bouncing between local minima, especially for those non-linearly separable pattern classification problems or complex function approximation problem. A second shortcoming is that the convergence of the algorithm is very sensitive to the initial value. So, it often converges to an inferior solution and gets trapped in a long training time.

4 Swarm Intelligence (SI)

Swarm Intelligence (SI) is a set of approaches based on the social insects, biological evolution, population based or biological behaviour of members in environment [22]. A common feature of population-based algorithms is that the population consisting of feasible solutions to the difficulty is customized by applying some agents on the solutions depending on the information of their robustness. Therefore, the population is encouraged towards improved solution areas within the solution space.

Since the last two decades, SI has been the focus of many researches because of its unique behaviour inherent of the social insects. He has defined the SI as “any attempt to design algorithm or distributed problem-solving devices inspired by the collective behaviour of social insect colonies and other animal societies." He mainly focused on the behaviour of social insects alone such as termites, bees, wasps, and different ant species [23].

5 Artificial Bee Colony Algorithm (ABC)

ABC was proposed for optimization, and NNs problem solution based on the intelligent foraging behaviour of honey bee swarm's behaviours [14, 24]. Therefore, ABC is more successful and most robust on multimodal functions included within the set with respect to DE, PSO, and GA [25]. The ABC algorithm provides a solution in the organized form by dividing the bee objects into different tasks such as employed, onlooker, and scout bees. These three bees/tasks determine the objects of problems by sharing information to other's bees.

Employed Bees: It used the multidirectional search space for food source with initialization of the area. They get information and all possibilities to find a food source and solution space. Sharing of information with onlooker bees is performed by employee bees. An employed bee produces a modification of the source position in her memory and discovers a new food source position. Provided that the nectar amount of the new source is higher than that of the previous source, the employed bee memorizes the new source position and forgets the old one. The e bee used the following equation (3) as.

$$v_{ij} = x_{ij} + \phi_{ij}\left(x_{ij} - x_{kj}\right) \tag{3}$$

Onlooker Bees: These bees evaluate the nectar amount obtained by employed bees and choose a food source depending on the probability values calculated using the fitness values. For this purpose, a fitness-based selection technique can be used. Onlooker bees watch the dance of hive bees and select the best food source according to the probability proportional to the quality of that food source. The following equation (4) shows the movement of onlooker’s bees for finding optimal food source.

$$v_{ij} = x_{ij} + \phi_{ij}\left(x_{ij} - x_{kj}\right) \tag{4}$$

Scout Bees: It selects the food source randomly without experience. If the nectar amount of a food source is higher than that of the previous source in their memory, they memorise the new position and forget the previous position. Whenever employed bees get a food source and use the food source very well again, they become scout bees to find new food sources by memorizing the best path by equation (5).

$$x_{ij}^{\mathrm{rand}} = x_{ij}^{\min} + \mathrm{rand}(0,1)\left(x_{ij}^{\max} - x_{ij}^{\min}\right) \tag{5}$$

6 Global Artificial Bee Colony (GABC) Algorithm

The ABC algorithm performance depends on bee agent such as employed, Onlooker's bees and Scout bees. Different techniques used to update the neighbour information equation to get better performance. SI based algorithms have the necessary properties of exploration and exploitation. In ABC, the exploration refers to the ability to explore the different unidentified sections in the solution space to discover the global optimum. Also the exploitation refers to the ability to apply the knowledge of the previous good solutions to find the best solution. Here GABC will collect the properties of exploration and exploitation with intelligent global behaviour of bees. GABC algorithm will update the solution step and will convert to the best solution on the basis of neighbourhood values. These modified steps will be in employed, onlooker and scout section.

Usually, in bee swarm, the experienced foragers can use previous knowledge of position and nectar quantity of food source to regulate their group directions in the search space. Furthermore, in ABC technique the best food can be finding through experience or neighbour cooperation. So GABC agents employed, scout and onlookers can be improved by their best food source. The GABC approach will merge their best finding approaches in standard ABC by the following steps.

Step 1: It modifies the employed section as

$$v_{ij} = x_{ij} + \phi_{ij}\left(x_{ij} - x_{kj}\right) + y \tag{6}$$

$$y = c_1 \mathrm{rand}(0,1)\left(x_j^{\mathrm{best}} - x_{ij}\right) + c_2 \mathrm{rand}(0,1)\left(y_j^{\mathrm{best}} - x_{ij}\right) \tag{7}$$

Step 2: Repeat the above formula with onlookers section

Where *y* shows Best_Food_Source, c_1 and c_2 are two constant values which is c_1 is 2.5 and c_2 is –3.5 for this study, x_j^{best} is the *j*-th element of the global best solution found so far, y_j^{best} is the *j*-th element of the best solution in the current iteration, ϕ_{ij} is a uniformly distributed real random number in the range $[-1,1]$.

Step 3: Modified the Scout section as

$$x_{ij}^{\mathrm{rand}} = x_{ij}^{\min} + \mathrm{rand}(0,1)\left(x_{ij}^{\max} - x_{ij}^{\min}\right) \tag{8}$$

If rand (0, 1) <=0.5, then

$$x_{ij}^{\mathrm{mutation}} = x_{ij} + \mathrm{rand}(0,1)\left(1 - \frac{iter}{iter_{\max}}\right)^{b} + \left(x_j^{best} - x_{ij}\right) \tag{9}$$

Else

$$x_{ij}^{\text{mutation}} = x_{ij} + \text{rand}\ (0,1)\left(1 - \frac{iter}{iter_{\max}}\right)^{b} + \left(y_{j}^{best} - x_{ij}\right) \qquad (10)$$

Then comparing the fitness value of random generated solution x_{ij}^{rand} and mutation solution x_{ij}^{mutation} the better one is chosen as a new food source, where b is a scaling parameter which is a positive integer within the range of [2,5].

7 Simulation Results

In this work, GABC algorithm used to train MLP for the boolean function classification task. To calculate the performance of the GABC, ABC and LM algorithms by Mean of Mean Square Error (MSE) and success rate for Boolean function classification, using Matlab 2010a software. The stopping criteria for minimum error set to 0.0001 LM while ABC and GABC stopped on MCN. During the experimentation, 30 trials performed in training. The sigmoid function used as activation function for network production. During the simulation, when the number of input signals, hidden nodes, output node and running time varies, performing training algorithms were stable, which is important for delegation NNs in the current state. The values of C_1 and C_2 were selected 2.5 and –3.5, respectively. From the simulation experiment, the GABC performance can be affected by C_1 and C_2. Here the best values selected for these two constant values.

Table 1. Parameters of the problems considered in the experiments

Problem	Colony Range	NN structure	D	MCN	Epoch	C1	C2
XOR13	[–10,10]	2–3–1+ Bias(4)	13	75	250	2.5	–3.5
3-Bit	[–10,10]	3–3–1+bias(4)	16	1000	1600	2.5	–3.5
Enc/Dec	[10,–10]	4–2–4+Bias(6)	22	1000	2100	2.5	–3.5

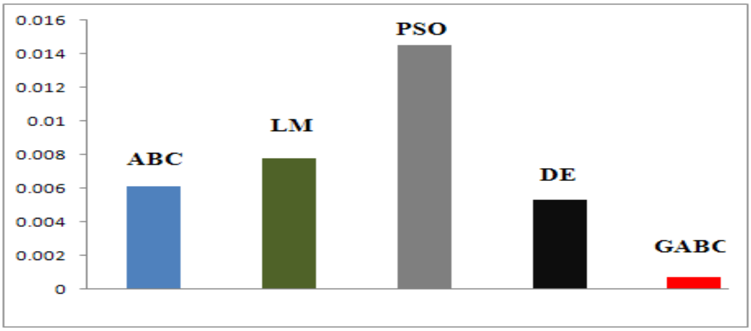


Fig. 2. Mean MSE of XOR13 problem using ABC, LM, PSO, DE and GABC

Where, D: Dimension of the problem, MCN: Maximum Cycle Numbers, NNs structure. From the following figures, the averages of MSE for XOR13, 3-bit Parity and 4-bit encoder/decoder are 0.000692, 0.00062, and 0.000294 respectively using GABC. The GABC have outstanding MSE for above problems from ABC and LM, where ABC is less than GABC for XOR6 only in term of MSE.

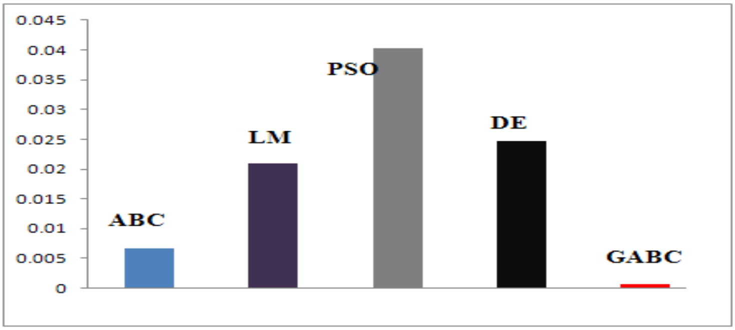


Fig. 3. Mean MSE of 3-Bit Parity Problem using ABC, LM, PSO, DE and GABC

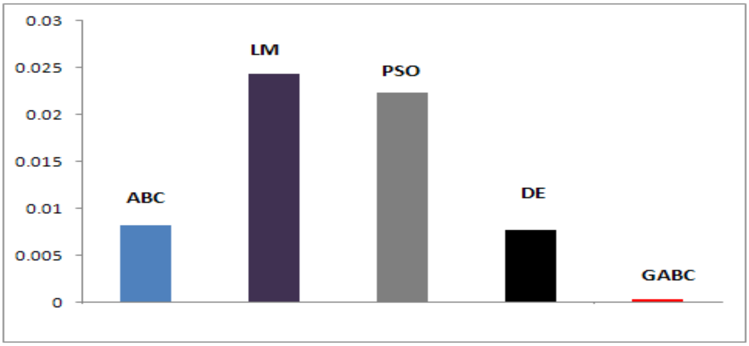


Fig. 4. Mean MSE of 4-Bit Encoder/Decoder using ABC, LM, PSO, DE and GABC

Table 2. Success Rate of ABC, BP (GD, LM) and GABCS algorithm

Problem Algorithm	XOR13	3-Bit Parity	4 Bit Dec/Encoder
ABC	100	100	100
LM	96.66	86.66	73.33
DE	100	77	100
PSO	63	93	80
GABCS	100	100	100

The testing Mean Square Error for XOR13, 3-bit Parity and 4-bit encoder/decoder are 0.0008,0067 and 0.00035 respectfully using GABC algorithm.

The success rate of GABC, ABC, PSO, DE and LM algorithms are given in table 4. The GABC has 100% success rate for XOR13, 3-bit Parity and 4-bit encoder or decoder problems, where GABC has 100% success rate.

8 Conclusion

The GABC algorithm collects the exploration and exploitation processes successfully by global method, which proves the high performance of the training MLP for Boolean classification. It has the best ability of searching global optimal solution.The proper weight may speed up the initialization and improve the classification accuracy.

Acknowledgement. The authors would like to thank University Tun Hussein Onn Malaysia for supporting this research under the FRGS Vote No 0739 and Postgraduate Incentive Research Grant Vote No 1032.

References

[1] Fionn, M.: Multilayer perceptrons for classification and regression. Neurocomputing 2, 183–197 (1991)

[2] Liao, S.-H., Wen, C.-H.: Artificial neural networks classification and clustering of methodologies and applications - literature analysis from 1995 to 2005. Expert Systems with Applications 32, 1–11 (2007)

[3] Ghazali, R., et al.: Non-stationary and stationary prediction of financial time series using dynamic ridge polynomial neural network. Neurocomputing 72, 2359–2367 (2009)

[4] Uncini, A.: Audio signal processing by neural networks. Neurocomputing 55, 593–625 (2003)

[5] Kiranyaz, S., et al.: Evolutionary artificial neural networks by multi-dimensional particle swarm optimization. Neural Networks 22, 1448–1462 (2009)

[6] Ilonen, J., et al.: Differential Evolution Training Algorithm for Feed-Forward Neural Networks. Neural Processing Letters 17, 93–105 (2003)

[7] Dorigo, M., Caro, G.D.: The Ant Colony Optimization Meta-Heuristic in New Ideas in Optimization. McGraw-Hill, England (1999)

[8] Kennedy, J., Eberhart, R.: Particle swarm optimization. In: Proceedings of IEEE International Conference on Neural Networks, vol. 4, pp. 1942–1948 (1995)

[9] Nawi, N.M., Ransing, R.S., Salleh, M.N.M., Ghazali, R., Hamid, N.A.: An Improved Back Propagation Neural Network Algorithm on Classification Problems. In: Zhang, Y., Cuzzocrea, A., Ma, J., Chung, K.-i., Arslan, T., Song, X. (eds.) DTA and BSBT 2010. CCIS, vol. 118, pp. 177–188. Springer, Heidelberg (2010)

[10] Nawi, N.M., Ghazali, R., Salleh, M.N.M.: The Development of Improved Back-Propagation Neural Networks Algorithm for Predicting Patients with Heart Disease. In: Zhu, R., Zhang, Y., Liu, B., Liu, C. (eds.) ICICA 2010. LNCS, vol. 6377, pp. 317–324. Springer, Heidelberg (2010)

[11] Rumelhart, D.E., et al.: Learning representations by back-propagating errors. Nature 323, 533–536 (1986)

[12] Shah, H., et al.: Global Hybrid Ant Bee Colony Algorithm for Training Artificial Neural Networks. Presented at the International Conference on Computational Science and Applications, Brazil (2012)

[13] Shah, H., Ghazali, R., Nawi, N.M.: Hybrid Ant Bee Colony Algorithm for Volcano Temperature Prediction. In: Chowdhry, B.S., Shaikh, F.K., Hussain, D.M.A., Uqaili, M.A. (eds.) IMTIC 2012. CCIS, vol. 281, pp. 453–465. Springer, Heidelberg (2012)

[14] Karaboga, D., Akay, B., Ozturk, C.: Artificial Bee Colony (ABC) Optimization Algorithm for Training Feed-Forward Neural Networks. In: Torra, V., Narukawa, Y., Yoshida, Y. (eds.) MDAI 2007. LNCS (LNAI), vol. 4617, pp. 318–329. Springer, Heidelberg (2007)

[15] Shah, H., Ghazali, R.: Prediction of Earthquake Magnitude by an Improved ABC-MLP. In: Developments in E-systems Engineering (DeSE), pp. 312–317 (2011)

[16] Peng, G., et al.: Global artificial bee colony search algorithm for numerical function optimization. In: 2011 Seventh International Conference on Natural Computation (ICNC), pp. 1280–1283 (2011)

[17] Shah, H., et al.: G-HABC Algorithm for Training Artificial Neural Networks. International Journal of Applied Metaheuristic Computing 3, 20 (2012)

[18] Stork, D.G., Allen, J.D.: How to solve the N-bit parity problem with two hidden units. Neural Networks 5, 923–926 (1992)

[19] Iyoda, E.M., et al.: A Solution for the N-bit Parity Problem Using a Single Translated Multiplicative Neuron. Neural Processing Letters 18, 233–238 (2003)

[20] Rosenblatt, F.: The Perceptron: A probabilistic model for information storage and organization in the brain. Psychological Review 65, 386–408 (1958)

[21] Rumelhart, D.E., et al.: Parallel distributed processing: Psychological and biological models. MIT Press (1986)

[22] Hieu Trung, H., Yonggwan, W.: Evolutionary algorithm for training compact single hidden layer feedforward neural networks. In: IEEE International Joint Conference on Neural Networks, IJCNN 2008, pp. 3028–3033 (2008)

[23] Bonabeau, E., et al.: Swarm Intelligence: From Natural to Artificial Systems. Oxford University Press, New York (1999)

[24] Karaboga, D., Akay, B.: A comparative study of Artificial Bee Colony algorithm. Applied Mathematics and Computation 214, 108–132 (2009)

[25] Karaboga, D., Kalinli, A.: Training recurrent neural networks for dynamic system identification using parallel tabu search algorithm. In: Proceedings of the 1997 IEEE International Symposium on Intelligent Control, pp. 113-118 (1997)

Fuzzy Decision Making Based on Hesitant Fuzzy Linguistic Term Sets

Li-Wei Lee[1] and Shyi-Ming Chen[2]

[1] Department of Computer and Communication Engineering, De Lin Institute of Technology,
New Taipei City, Taiwan, R.O.C.
[2] Department of Computer Science and Information Engineering,
National Taiwan University of Science and Technology, Taipei, Taiwan, R.O.C.

Abstract. This paper presents a new fuzzy decision making method based on likelihood-based comparison relations of hesitant fuzzy linguistic term sets. We also present a similarity measure between hesitant fuzzy linguistic term sets. The proposed method is more simple for fuzzy decision making than the method presented in [18]. It provides us with a useful way for decision making in a fuzzy environment.

Keywords: Hesitant fuzzy sets, hesitant fuzzy linguistic term sets, fuzzy decision making, likelihood-based comparison relations, similarity measures.

1 Introduction

The fuzzy linguistic approach has successfully been applied to deal with decision making problems [3]-[19], [21]-[23]. In a fuzzy decision making environment, experts maybe hesitate to choose appropriate linguistic terms to assess alternatives in some situations for reaching a final agreement. In order to deal with such situations, Torra [20] presented the concept of hesitant fuzzy sets, which is a generalization of fuzzy sets [26]. He also presented different generalizations and extensions of fuzzy sets and discussed the relationships among hesitant fuzzy sets and the other generalizations of fuzzy sets, such as intuitionistic fuzzy sets [1], [2], type 2 fuzzy sets [8], [15], type n fuzzy sets [8] and fuzzy multisets [16]. Based on the concept of hesitant fuzzy sets presented in [20], some researchers [18], [25], [27] have studied related issues of hesitant fuzzy sets. In [18], Rodriguez et al. presented the concept of hesitant fuzzy linguistic term sets based on the fuzzy linguistic approach [26] and hesitant fuzzy sets [20]. They pointed out that the fuzzy linguistic approach is very limited due to the fact that it assesses a linguistic variable by using a single linguistic term, whereas the hesitant fuzzy linguistic term sets approach assesses a linguistic variable by using several linguistic terms for decision making. They presented two symbolic aggregation operators to obtain a linguistic interval associated with each alternative and presented an exploitation process to get a preference order for decision making based on the nondominance choice degree of a preference relation obtained from linguistic intervals. However, the drawback of the method presented in [18] is that it is too complicated for dealing with fuzzy decision making problems. Therefore, we

A. Selamat et al. (Eds.): ACIIDS 2013, Part I, LNAI 7802, pp. 21–30, 2013.

must develop a new fuzzy decision making method based on hesitant fuzzy linguistic term sets to overcome the drawback of the method presented in [18].

In this paper, we present a new fuzzy decision making method based on likelihood-based comparison relations of hesitant fuzzy linguistic term sets. We also present a similarity measure between hesitant fuzzy linguistic term sets. The proposed fuzzy decision making method is more simple for fuzzy decision making than the method presented in [18].

The rest of this paper is organized as follows. In Section 2, we briefly review the concept of hesitant fuzzy linguistic term sets [18]. In Section 3, we present the concept of likelihood-based comparison relations and present a similarity measure of hesitant fuzzy linguistic term sets. In Section 4, we present a fuzzy decision making method based on likelihood-based comparison relations of hesitant fuzzy linguistic term sets. The conclusions are discussed in Section 5.

2 A Review of Rodriguez et al.'s Decision Making Method Based on Hesitant Fuzzy Linguistic Term Sets

In [18], Rodriguez et al. presented the concept of hesitant fuzzy linguistic term sets for decision making. The basic concepts and operations of hesitant fuzzy linguistic term sets are reviewed from [18] as follows.

***Definition 2.1* [18]:** Let $S = \{s_0, s_1, \ldots, s_g\}$ be a linguistic term set. A hesitant fuzzy linguistic term set H_S is an ordered finite subset of consecutive linguistic terms of the linguistic term set S.

***Definition 2.2* [18]:** Let $G_H = (V_N, V_T, I, P)$ be a context-free grammar and let $S = \{s_0, s_1, \ldots, s_g\}$ be a linguistic term set, where V_N denotes a set of nonterminal symbols, V_T denotes a set of terminal symbols, I denotes the starting symbol and P denotes the production rules, shown as follows:

$$
\begin{aligned}
&V_N = \{\langle\text{primary term}\rangle, \langle\text{composite term}\rangle, \langle\text{unary relation}\rangle, \langle\text{binary relation}\rangle,\\
&\qquad \langle\text{conjunction}\rangle\},\\
&V_T = \{\text{lower than, greater than, between, } s_0, s_1, \ldots, \text{ and } s_g\},\\
&I \in V_N,\\
&P = \{I ::= \langle\text{primary term}\rangle | \langle\text{composite term}\rangle,\\
&\quad \langle\text{composite term}\rangle ::= \langle\text{unary relation}\rangle\langle\text{primary term}\rangle | \langle\text{binary relation}\rangle\\
&\qquad\qquad \langle\text{primary term}\rangle\langle\text{conjunction}\rangle\langle\text{primary term}\rangle,\\
&\quad \langle\text{primary term}\rangle ::= s_0 | s_1 | \cdots | s_g,\\
&\quad \langle\text{unary relation}\rangle ::= \text{lower than} | \text{greater than},\\
&\quad \langle\text{binary relation}\rangle ::= \text{between},\\
&\quad \langle\text{conjunction}\rangle ::= \text{and}\}.
\end{aligned}
$$

***Definition 2.3* [18]:** Let E_{G_H}be a function that transforms the linguistic expressions le obtained by the context-free grammar G_H into a hesitant fuzzy linguistic term set $H_S$of the linguistic term set S, shown as follows:

$$E_{G_H}: le \to H_S.$$

The linguistic expressions generated by production rules can be transformed into a hesitant fuzzy linguistic term set in different ways according to their meaning:

1) $E_{G_H}(s_i) = \{s_i | s_i \in S\}$,
2) $E_{G_H}(\text{less than} s_i) = \{s_j | s_j \in S \text{ and } s_j \le s_i\}$,
3) $E_{G_H}(\text{greater than} s_i) = \{s_j | s_j \in S \text{ and } s_j \ge s_i\}$,
4) $E_{G_H}(\text{between } s_i \text{and } s_j) = \{s_k | s_k \in S \text{ and } s_i \le s_k \le s_j\}$.

Definition 2.4 **[18]:** Let $X = \{x_1, x_2, \ldots, x_n\}$ be a set of alternatives, let $C = \{c_1, c_2, \ldots, c_m\}$ be a set of criteria, let $S = \{s_0, s_1, \ldots, s_g\}$ be a linguistic term set and let $H_S^j(x_i)$ be a hesitant fuzzy linguistic term set associated with alternative x_i with respect to criterion c_j, where $1 \le i \le n$ and $1 \le j \le m$. The min_upper operator $H_{S_{min}^+}(x_i)$ and the max_lower operator $H_{S_{max}^-}(x_i)$ of alternative x_i are defined as follows:

$$H_{S_{min}^+}(x_i) = \min\{H_{S^+}^j(x_i) | \, 1 \le i \le n \text{ and } 1 \le j \le m\}, \tag{1}$$

$$H_{S_{max}^-}(x_i) = \max\{H_{S^-}^j(x_i) | \, 1 \le i \le n \text{ and } 1 \le j \le m\}, \tag{2}$$

where $H_{S^+}^j(x_i)$ and $H_{S^-}^j(x_i)$ are the upper bound of hesitant fuzzy linguistic term set $H_S^j(x_i)$ and the lower bound of hesitant fuzzy linguistic term set $H_S^j(x_i)$ associated with alternative x_i with respect to criterion c_j, respectively.

Based on the min_upper operator $H_{S_{min}^+}(x_i)$ and the max_lower operator $H_{S_{max}^-}(x_i)$, the linguistic interval $H'(x_i)$ for each alternative x_i can be obtained, shown as follows [18]:

$$H'(x_i) = \left[min\left\{H_{S_{min}^+}(x_i), H_{S_{max}^-}(x_i)\right\}, max\left\{H_{S_{min}^+}(x_i), H_{S_{max}^-}(x_i)\right\}\right]. \tag{3}$$

Definition 2.5 **[18]:** Let P be a preference relation defined over a set X of alternatives. For alternative x_i, its nondominance choice degree NDD_i is obtained as follows:

$$NDD_i = \min\{1 - p_{ji}^S, j \ne i\}, \tag{4}$$

where $p_{ij} = p(x_i \ge x_j) = \frac{\max(0, x_{iR} - x_{jL}) - \max(0, x_{iL} - x_{jR})}{(x_{iR} - x_{iL}) + (x_{jR} - x_{jL})}$ denotes the degree of the alternative x_i over x_j, $1 \le i \le n$, $1 \le j \le n$, $i \ne j$, $x_i = [x_{iL}, x_{iR}]$, $x_j = [x_{jL}, x_{jR}]$, and $p_{ji}^S = \max\{p_{ji} - p_{ij}, 0\}$ represents the degree in which x_i is strictly dominated by x_j. The larger the value of NDD_i, the better the preference order of alternative x_i, where $1 \le i \le n$.

3 Likelihood-Based Comparison Relations and Similarity Measures of Hesitant Fuzzy Linguistic Term Sets

In this section, we propose the concept of likelihood-based comparison relations of hesitant fuzzy linguistic term sets and propose a similarity measure between hesitant fuzzy linguistic term sets. Assume that there is a linguistic term set $S = \{s_0, s_1, \ldots, s_g\}$, where the membership functions of the linguistic terms in linguistic term set S are shown in Fig. 1.

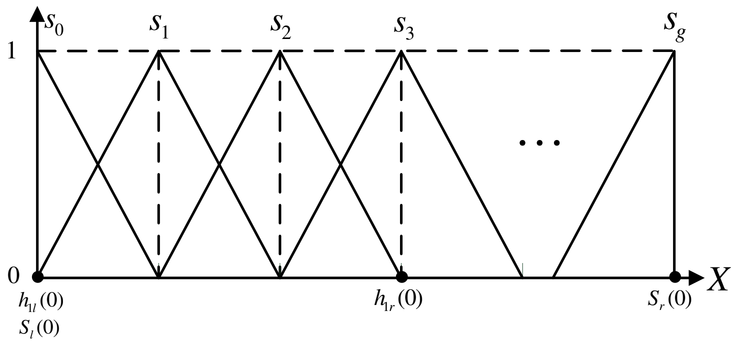


Fig. 1. The 0-cut $h_1(0)$ of the hesitant fuzzy linguistic term set $h_1 = \{s_0, s_1, s_2\}$ and the 0-cut $S(0)$ of the linguistic term set $S = \{s_0, s_1, \ldots, s_g\}$

Assume that there is a hesitant fuzzy linguistic term set $h_1 = \{s_0, s_1, s_2\}$, then the 0-cut $h_1(0)$ of the hesitant fuzzy linguistic term set h_1 is defined as follows:

$$h_1(0) = [h_{1l}(0), h_{1r}(0)],$$

where the 0-cut $h_1(0)$ of h_1 and the 0-cut $S(0)$ of S are shown in Fig. 1, respectively, where $S(0) = [S_l(0), S_r(0)]$. From Fig. 1, we can see that $h_{1l}(0)$ has the largest membership degree in the membership function of the linguistic term s_0 and $h_{1r}(0)$ has the largest membership degree in the membership function of the linguistic term s_3. Therefore, we can get $h_1(0) = [s_{0,}s_3]$. From Fig. 1, we also can see that $S_l(0)$ has the largest membership degree in the membership function of the linguistic term s_0 and $S_r(0)$ has the largest membership degree in the membership function of the linguistic term s_g. Therefore, we can get $S(0) = [s_{0,}s_g]$.

Definition 3.1: Assume that there is a linguistic term set $S = \{s_0, s_1, \ldots, s_g\}$. Based on the concept of likelihood-based comparison relations between intervals [24], the likelihood-based comparison relation $p(h_1 \geq h_2)$ between two hesitant fuzzy linguistic term sets h_1 and h_2 is defined as follows:

$$p(h_1 \geq h_2) = \max\left(1 - \max\left(\frac{\mathrm{Ind}(h_{2r}(0)) - \mathrm{Ind}(h_{1l}(0))}{L(h_1(0)) + L(h_2(0))}, 0\right), 0\right), \tag{5}$$

where $h_1(0) = [h_{1l}(0), h_{1r}(0)]$ is the 0-cut of the hesitant fuzzy linguistic term set h_1, $h_2(0) = [h_{2l}(0), h_{2r}(0)]$ is the 0-cut of the hesitant fuzzy linguistic term set h_2, $\mathrm{Ind}(s_i) = i$ denotes the index associated with the linguistic term s_i, $L(h_1(0)) = \mathrm{Ind}(h_{1r}(0)) - \mathrm{Ind}(h_{1l}(0))$ and $L(h_2(0)) = \mathrm{Ind}(h_{2r}(0)) - \mathrm{Ind}(h_{2l}(0))$.

The likelihood-based comparison relation $p(h_1 \geq h_2)$ between two hesitant fuzzy linguistic term sets h_1 and h_2 has the following properties:

1) $0 \leq p(h_1 \geq h_2) \leq 1$.
2) $p(h_1 \geq h_2) + p(h_2 \geq h_1) = 1$.
3) If $\mathrm{Ind}\big(h_{1r}(0)\big) \leq \mathrm{Ind}\big(h_{2l}(0)\big)$, then $p(h_1 \geq h_2) = 0$.

4) If $\mathrm{Ind}(h_{1l}(0)) \geq \mathrm{Ind}(h_{2r}(0))$, then $p(h_1 \geq h_2) = 1$.
5) $p(h_1 \geq h_1) = 0.5$.

If $\mathrm{Ind}(h_{1r}(0)) = \mathrm{Ind}(h_{1l}(0))$ and $\mathrm{Ind}(h_{2r}(0)) = \mathrm{Ind}(h_{2l}(0))$, then the likelihood-based comparison relation $p(h_1 \geq h_2)$between two hesitant fuzzy linguistic term sets h_1 and h_2 is defined as follows:

$$p(h_1 \geq h_2) = \begin{cases} 1, \mathrm{if} \mathrm{Ind}(h_{1l}(0)) > \mathrm{Ind}(h_{2l}(0)) \\ \frac{1}{2}, \mathrm{if} \mathrm{Ind}(h_{1l}(0)) = \mathrm{Ind}(h_{2l}(0)) \\ 0, \mathrm{if} \mathrm{Ind}(h_{1l}(0)) < \mathrm{Ind}(h_{2l}(0)) \end{cases}$$

Definition 3.2: Assume that there is a linguistic term set $S = \{s_0, s_1, \ldots, s_g\}$. The degree of similarity $s(h_1, h_2)$ between two hesitant fuzzy linguistic term sets h_1 and h_2 is defined as follows:

$$s(h_1, h_2) = 1 - |p(h_1 \geq S) - p(h_2 \geq S)|. \tag{6}$$

4 The Proposed Fuzzy Decision Making Method Based on Likelihood-Based Comparison Relations of Hesitant Fuzzy Linguistic Term Sets

In this section, we present a fuzzy decision making method based on likelihood-based comparison relations of hesitant fuzzy linguistic term sets. Assume that there is a set X of alternatives, $X = \{x_1, x_2, \ldots, x_n\}$, assume that there is a set C of criteria, $C = \{c_1, c_2, \ldots, c_m\}$, and assume that there is a set W of weights, $W = \{\omega_1, \omega_2, \ldots, \omega_m\}$, where ω_j denotes the weight of criterion c_j and $1 \leq j \leq m$. Assume that there is a linguistic term set $S = \{s_0, s_1, \ldots, s_g\}$ and assume that there is a context-free grammar G_H which produces the linguistic expressions of alternatives with respect to different criteria, where the linguistic expressions are transformed into hesitant fuzzy linguistic term sets by means of the transformation function E_{G_H}. Based on the proposed likelihood-based comparison relations of hesitant fuzzy linguistic term sets, the proposed fuzzy decision making method is now presented as follows:

Step 1: Construct the decision matrix Y, shown as follows:

$$Y = \begin{array}{c} \\ x_1 \\ x_2 \\ \vdots \\ x_n \end{array} \begin{array}{c} \begin{array}{cccc} c_1 & c_2 & \cdots & c_m \end{array} \\ \begin{bmatrix} y_{11} & y_{12} & \cdots & y_{1m} \\ y_{21} & y_{22} & \cdots & y_{2m} \\ \vdots & \vdots & \vdots & \vdots \\ y_{n1} & y_{n2} & \cdots & y_{nm} \end{bmatrix} \end{array},$$

where y_{ij} is a hesitant fuzzy linguistic term set of alternative x_i with respect to criterion c_j, $1 \leq i \leq n$ and $1 \leq j \leq m$.

Step 2: Based on Eq. (5), construct the likelihood-based comparison relation *P*, shown as follows:

$$P = \begin{array}{c} \\ x_1 \\ x_2 \\ \vdots \\ x_n \end{array} \begin{array}{c} \begin{array}{cccc} c_1 & c_2 & \cdots & c_m \end{array} \\ \begin{bmatrix} p_{11} & p_{12} & \cdots & p_{1m} \\ p_{21} & p_{22} & \cdots & p_{2m} \\ \vdots & \vdots & \vdots & \vdots \\ p_{n1} & p_{n2} & \cdots & p_{nm} \end{bmatrix} \end{array},$$

$$p_{ij} = p\left(y_{ij} \geq S\right) = \max\left(1 - \max\left(\frac{\mathrm{Ind}(S_r(0)) - \mathrm{Ind}(y_{ijl}(0))}{L(y_{ij}(0)) + L(S(0))}, 0\right), 0\right), \tag{7}$$

where $y_{ij}(0) = [y_{ijl}(0), y_{ijr}(0)]$ is the 0-cut of the hesitant fuzzy linguistic term set y_{ij}, $S(0) = [S_l(0), S_r(0)]$ is the 0-cut of the linguistic term set S, $1 \leq i \leq n$ and $1 \leq j \leq m$.

Step 3: Let

$$R(x_i) = \sum_{j=1}^{m} \omega_j p_{ij}, \tag{8}$$

where ω_j denotes the weight of criterion c_j, $1 \leq i \leq n$ and $1 \leq j \leq m$. The larger the value of $R(X_i)$, the better the preference order of alternative x_i, where $1 \leq i \leq n$.

In the following, we use an example to illustrate the process of the proposed fuzzy decision making method.

Example 4.1:Assume that there are three alternatives x_1, x_2, x_3, assume that there are three criteria c_1, c_2, c_3, and assume that the weights of the criteria c_1, c_2 and c_3 are 1/3, 1/3 and 1/3, respectively. Assume that there is a linguistic term set S = {s_0: nothing (*n*), s_1: very low (*vl*), s_2: low (*l*), s_3: medium (*m*), s_4: high (*h*), s_5: very high (*vh*), s_6: perfect (*p*)}. Thelinguistic expressions of the alternatives with respect to different criteria are shown in Table 1. Based on the transformation function E_{G_H} shown in *Definition 2.3*, Table 1 can be transformed into Table 2.

Table 1. Linguistic expressions of the alternatives with respect to different criteria [18]

Criteria Alternatives	c_1	c_2	c_3
x_1	between *vl* and *m*	between *h* and *vh*	*h*
x_2	between *l* and *m*	*m*	lower than *l*
x_3	greater than *h*	between *vl* and *l*	greater than *h*

Table 2. Transformation of Table 1 into hesitant fuzzy linguistic term sets [18]

Criteria Alternatives	c_1	c_2	c_3
x_1	{*vl*, *l*, *m*}	{*h*, *vh*}	{*h*}
x_2	{*l*, *m*}	{*m*}	{*n*, *vl*, *l*}
x_3	{*h*, *vh*, *p*}	{*vl*, *l*}	{*h*, *vh*, *p*}

The fuzzy decision making process based on the proposed method is shown as follows:

[Step 1]: We can get the decision matrix *Y*, shown as follows:

$$Y = \begin{array}{cc} & \begin{array}{ccc} \quad c_1 \quad & \quad c_2 \quad & \quad c_3 \quad \end{array} \\ \begin{array}{c} x_1 \\ x_2 \\ x_3 \end{array} & \begin{bmatrix} \{vl, l, m\} & \{h, vh\} & \{h\} \\ \{l, m\} & \{m\} & \{n, vl, l\} \\ \{h, vh, p\} & \{vl, l\} & \{h, vh, p\} \end{bmatrix} \end{array}.$$

[Step 2]: Based on Eq. (7), we can get the likelihood-based comparison relation *P*, shown as follows:

$S(0) = [s_0, s_6], y_{11}(0) = [s_0, s_4], y_{12}(0) = [s_3, s_6], y_{13}(0) = [s_3, s_5], y_{21}(0) = [s_1, s_4], y_{22}(0) = [s_2, s_4], y_{23}(0) = [s_0, s_3], y_{31}(0) = [s_3, s_6], y_{32}(0) = [s_0, s_3], y_{33}(0) = [s_3, s_6],$

$$p_{11} = p(y_{11} \geq S) = \max\left(1 - \max\left(\frac{\operatorname{Ind}(S_r(0)) - \operatorname{Ind}(y_{11l}(0))}{L(y_{11}(0)) + L(S(0))}, 0\right), 0\right),$$
$$= \max\left(1 - \max\left(\frac{\operatorname{Ind}(s_6) - \operatorname{Ind}(s_0)}{\operatorname{Ind}(s_4) - \operatorname{Ind}(s_0) + \operatorname{Ind}(s_6) - \operatorname{Ind}(s_0)}, 0\right), 0\right)$$
$$= 0.4000,$$

$$p_{12} = p(y_{12} \geq S) = \max\left(1 - \max\left(\frac{\operatorname{Ind}(S_r(0)) - \operatorname{Ind}(y_{12l}(0))}{L(y_{12}(0)) + L(S(0))}, 0\right), 0\right),$$
$$= \max\left(1 - \max\left(\frac{\operatorname{Ind}(s_6) - \operatorname{Ind}(s_3)}{\operatorname{Ind}(s_6) - \operatorname{Ind}(s_3) + \operatorname{Ind}(s_6) - \operatorname{Ind}(s_0)}, 0\right), 0\right)$$
$$= 0.6667,$$

$$p_{13} = p(y_{13} \geq S) = \max\left(1 - \max\left(\frac{\operatorname{Ind}(S_r(0)) - \operatorname{Ind}(y_{13l}(0))}{L(y_{13}(0)) + L(S(0))}, 0\right), 0\right),$$
$$= \max\left(1 - \max\left(\frac{\operatorname{Ind}(s_6) - \operatorname{Ind}(s_3)}{\operatorname{Ind}(s_5) - \operatorname{Ind}(s_3) + \operatorname{Ind}(s_6) - \operatorname{Ind}(s_0)}, 0\right), 0\right)$$
$$= 0.6250,$$

$$p_{21} = p(y_{21} \geq S) = \max\left(1 - \max\left(\frac{\operatorname{Ind}(S_r(0)) - \operatorname{Ind}(y_{21l}(0))}{L(y_{21}(0)) + L(S(0))}, 0\right), 0\right),$$
$$= \max\left(1 - \max\left(\frac{\operatorname{Ind}(s_6) - \operatorname{Ind}(s_1)}{\operatorname{Ind}(s_4) - \operatorname{Ind}(s_1) + \operatorname{Ind}(s_6) - \operatorname{Ind}(s_0)}, 0\right), 0\right)$$
$$= 0.4444,$$

$$p_{22} = p(y_{22} \geq S) = \max\left(1 - \max\left(\frac{\operatorname{Ind}(S_r(0)) - \operatorname{Ind}(y_{22l}(0))}{L(y_{22}(0)) + L(S(0))}, 0\right), 0\right),$$
$$= \max\left(1 - \max\left(\frac{\operatorname{Ind}(s_6) - \operatorname{Ind}(s_2)}{\operatorname{Ind}(s_4) - \operatorname{Ind}(s_2) + \operatorname{Ind}(s_6) - \operatorname{Ind}(s_0)}, 0\right), 0\right)$$
$$= 0.5000,$$

$$p_{23} = p(y_{23} \geq S) = \max\left(1 - \max\left(\frac{\operatorname{Ind}(S_r(0)) - \operatorname{Ind}(y_{23l}(0))}{L(y_{23}(0)) + L(S(0))}, 0\right), 0\right),$$
$$= \max\left(1 - \max\left(\frac{\operatorname{Ind}(s_6) - \operatorname{Ind}(s_0)}{\operatorname{Ind}(s_3) - \operatorname{Ind}(s_0) + \operatorname{Ind}(s_6) - \operatorname{Ind}(s_0)}, 0\right), 0\right)$$
$$= 0.3333,$$

$$
\begin{aligned}
p_{31} &= p(y_{31} \geq S) = \max\left(1 - \max\left(\frac{\mathrm{Ind}(S_r(0)) - \mathrm{Ind}(y_{31l}(0))}{L(y_{31}(0)) + L(S(0))}, 0\right), 0\right), \\
&= \max\left(1 - \max\left(\frac{\mathrm{Ind}(s_6) - \mathrm{Ind}(s_3)}{\mathrm{Ind}(s_6) - \mathrm{Ind}(s_3) + \mathrm{Ind}(s_6) - \mathrm{Ind}(s_0)}, 0\right), 0\right) \\
&= 0.6667, \\
p_{32} &= p(y_{32} \geq S) = \max\left(1 - \max\left(\frac{\mathrm{Ind}(S_r(0)) - \mathrm{Ind}(y_{32l}(0))}{L(y_{32}(0)) + L(S(0))}, 0\right), 0\right), \\
&= \max\left(1 - \max\left(\frac{\mathrm{Ind}(s_6) - \mathrm{Ind}(s_0)}{\mathrm{Ind}(s_3) - \mathrm{Ind}(s_0) + \mathrm{Ind}(s_6) - \mathrm{Ind}(s_0)}, 0\right), 0\right) \\
&= 0.3333, \\
p_{33} &= p(y_{33} \geq S) = \max\left(1 - \max\left(\frac{\mathrm{Ind}(S_r(0)) - \mathrm{Ind}(y_{33l}(0))}{L(y_{33}(0)) + L(S(0))}, 0\right), 0\right), \\
&= \max\left(1 - \max\left(\frac{\mathrm{Ind}(s_6) - \mathrm{Ind}(s_3)}{\mathrm{Ind}(s_6) - \mathrm{Ind}(s_3) + \mathrm{Ind}(s_6) - \mathrm{Ind}(s_0)}, 0\right), 0\right) \\
&= 0.6667,
\end{aligned}
$$

$$
P = \begin{array}{c} \\ x_1 \\ x_2 \\ x_3 \end{array}
\begin{array}{c} \begin{array}{ccc} c_1 & c_2 & c_3 \end{array} \\
\begin{bmatrix} 0.4000 & 0.6667 & 0.6250 \\ 0.4444 & 0.5000 & 0.3333 \\ 0.6667 & 0.3333 & 0.6667 \end{bmatrix} \end{array}.
$$

[Step 3]: Because the weights $\omega_1, \omega_2, \omega_3$ of the criterion c_1, c_2, c_3 are 1/3, 1/3 and 1/3, respectively, based on Eq. (8), we can get

$$R(x_1) = \sum_{j=1}^{3} \omega_j p_{1j} = \frac{1}{3} \times 0.4000 + \frac{1}{3} \times 0.6667 + \frac{1}{3} \times 0.6250 = 0.5639,$$

$$R(x_2) = \sum_{j=1}^{3} \omega_j p_{2j} = \frac{1}{3} \times 0.4444 + \frac{1}{3} \times 0.5000 + \frac{1}{3} \times 0.3333 = 0.4259,$$

$$R(x_3) = \sum_{j=1}^{3} \omega_j p_{3j} = \frac{1}{3} \times 0.6667 + \frac{1}{3} \times 0.3333 + \frac{1}{3} \times 0.6667 = 0.5556.$$

Because $R(x_1) > R(x_3) > R(x_2)$, the preference order of the alternatives x_1, x_2 and x_3 is: $x_1 > x_3 > x_2$.. This result coincides with the one presented in [18].

5 Conclusions

We have presented the concept of likelihood-based comparison relations of hesitant fuzzy linguistic term sets. We also have presented a similarity measure between hesitant fuzzy linguistic term sets. Based on likelihood-based comparison relations of hesitant fuzzy linguistic term sets, we have presented a new method for fuzzy decision making. The proposed method is more simple for fuzzy decision making than the method presented in [18].

Acknowledgements. This work was supported in part by the Taiwan Ubiquitous Information Co., Ltd, Kaohsiung, Taiwan, Republic of China, and was supported in part by the National Science Council, Republic of China, under Grant NSC 100-2221-E-011-118-MY2.

References

1. Atanassov, K.: Intuitionistic Fuzzy Sets. Fuzzy Sets and Systems 20, 87–96 (1986)
2. Atanassov, K., Gargov, G.: Interval Valued Intuitionistic Fuzzy Sets. Fuzzy Sets and Systems 31, 343–349 (1989)
3. Ben-Arieh, D., Chen, Z.: Linguistic-Labels Aggregation and Consensus Measure for Autocratic Decision Making Using Group Recommendations. IEEE Transactions on Systems, Man, and Cybernetics-Part A: Systems and Humans 36, 558–568 (2006)
4. Chen, S.M., Lee, L.W.: Autocratic Decision Making Using Group Recommendations Based on the ILLOWA Operators and Likelihood-Based Comparison Relations. IEEE Transactions on Systems, Man, and Cybernetics-Part A: Systems and Humans 42, 115–129 (2012)
5. Chen, S.M., Lee, L.W., Yang, S.W., Sheu, T.W.: Adaptive Consensus Support Model for Group Decision Making Systems. Expert Systems with Applications 39, 12580–12588 (2012)
6. Chen, S.M., Lee, L.W., Liu, H.C., Yang, S.W.: Multiattribute Decision Making Based on Interval-Valued Intuitionistic Fuzzy Values. Expert Systems with Applications 39, 10343–10351 (2012)
7. Dong, Y., Xu, Y., Yu, S.: Computing the Numerical Scale of the Linguistic Term Set for the 2-Tuple Fuzzy Linguistic Representation Model. IEEE Transactions on Fuzzy Systems 17, 1366–1378 (2009)
8. Dubois, D., Prade, H.: Fuzzy Sets and Systems: Theory and Applications. Kluwer, New York (1980)
9. Herrera, F., Herrera-Viedma, E., Martinez, L.: A Fuzzy Linguistic Methodology to Deal with Unbalanced Linguistic Term Sets. IEEE Transactions on Fuzzy Systems 16, 354–370 (2008)
10. Herrera, F., Martinez, L.: A 2-Tuple Fuzzy Linguistic Representation Model for Computing with Words. IEEE Transactions on Fuzzy Systems 8, 746–752 (2000)
11. Herrera-Viedma, E., Martinez, L., Mata, F., Chiclana, F.: A Consensus Support System Model for Group Decision-Making Problems with Multi-Granular Linguistic Preference Relations. IEEE Transactions on Fuzzy Systems 13, 644–658 (2005)
12. Ma, J., Ruan, D., Xu, Y., Zhang, G.: A Fuzzy-Set Approach to Treat Determinacy and Consistency of Linguistic Terms in Multi-Criteria Decision Making. International Journal of Approximate Reasoning 44, 165–181 (2007)
13. Martinez, L.: Sensory Evaluation Based on Linguistic Decision Analysis. International Journal of Approximate Reasoning 44, 148–164 (2007)
14. Mata, F., Martinez, L., Herrera-Viedma, E.: An Adaptive Consensus Support Model for Group Decision-Making Problems in A Multi-Granular Fuzzy Linguistic Context. IEEE Transactions on Fuzzy Systems 17, 279–290 (2009)
15. Mendel, J.M., John, R.I.B.: Type-2 Fuzzy Sets Made Simple. IEEE Transactions on Fuzzy Systems 10, 117–127 (2002)
16. Miyamoto, S.: Generalization of Multisets and Rough Approximations. International Journal of Intelligent Systems 19, 639–652 (2004)
17. Pedrycz, W., Song, M.: Analytic Hierarchy Process (AHP) in Group Decision Making and Its Optimization with An Allocation of Information Granularity. IEEE Transactions on Fuzzy Systems 19, 527–539 (2011)
18. Rodriguez, R.M., Martinez, L., Herrera, F.: Hesitant Fuzzy Linguistic Term Sets for Decision Making. IEEE Transactions on Fuzzy Systems 20, 109–119 (2012)
19. Tang, Y., Zheng, J.: Linguistic Modeling Based on Semantic Similarity Relation among Linguistic Labels. Fuzzy Sets and Systems 157, 1662–1673 (2006)

20. Torra, V.: Hesitant Fuzzy Sets. International Journal of Intelligent Systems 25, 529–539 (2010)
21. Wang, J.H., Hao, J.: A New Version of 2-Tuple Fuzzy Linguistic Representation Model for Computing with Words. IEEE Transactions on Fuzzy Systems 14, 435–445 (2006)
22. Wu, D., Mendel, J.M.: Aggregation Using the Linguistic Weighted Average and Interval Type-2 Fuzzy Sets. IEEE Transactions on Fuzzy Systems 15, 1145–1161 (2007)
23. Wu, D., Mendel, J.M.: Computing with Words for Hierarchical Decision Making Applied to Evaluating A Weapon System. IEEE Transactions on Fuzzy Systems 18, 441–460 (2010)
24. Xu, Z.S., Da, Q.L.: A Likelihood-Based Method for Priorities of Interval Judgment Matrices. Chinese Journal of Management Science 11, 63–65 (2003)
25. Xu, Z., Xia, M.: Distance and Similarity Measures for Hesitant Fuzzy Sets. Information Sciences 181, 2128–2138 (2011)
26. Zadeh, L.A.: Fuzzy Sets. Information and Control 8, 338–353 (1965)
27. Zhu, B., Xu, Z., Xia, M.: Hesitant Fuzzy Geometric Bonferroni Means. Information Sciences 205, 72–85 (2012)

Time-Varying Mutation in Particle Swarm Optimization

S. Masrom, Siti. Z.Z. Abidin, N. Omar, and K. Nasir

Faculty of Computer and Mathematical Sciences
Universiti Teknologi MARA, Shah Alam, Selangor, Malaysia
suray078@perak.uitm.edu.my,
sitizaleha533@salam.uitm.edu.my,
nasiroh@tmsk.uitm.edu.my,
kalsom@tmsk.uitm.edu.my

Abstract. One of significant improvement for particle swarm optimization (PSO) is through the implementation of metaheuristics hybridization that combines different metaheuristics paradigms. By using metaheuristics hybridization, the weaknesses of one algorithm can be compensated by the strengths of other algorithms. Therefore, researchers have given a lot of interest in hybridizing PSO with mutation concept from genetic algorithm (GA). The reason for incorporating mutation into PSO is to resolve premature convergence problem due to some kind of stagnation by PSO particles. Although PSO is capable to produce fast results, particles stagnation has led the algorithm to suffer from low-optimization precision. Thus, this paper introduces time-varying mutation techniques for resolving the PSO problem. The different time-varying techniques have been tested on some benchmark functions. Results from the empirical experiments have shown that most of the time-varying mutation techniques have significantly improved PSO performances not just to the results accuracy but also to the convergence time.

Keywords: Hybridization, Particle swarm optimization, Time-varying, Mutation, Inertia weight.

1 Introduction

Inspired from the cognitive and social behavior of bird flocking and fish schooling, a kind of metaheuristic algorithm has been proposed by Kennedy and Ebehart in 1995 named Particle Swarm Optimization (PSO)[1, 2]. This algorithm has received wide attention by many researchers in solving many kinds of optimization problems including scheduling, routing, resource allocating and time tabling.

Nevertheless, the optimization problems are very computationally expensive. In other words, the time taken for completing the search algorithm is exponential to the problem size. Although PSO is very effective in providing fast results, its ability in achieving high accurate solutions especially for real life problems is still insufficient [3]. Besides, as a meteheuristic algorithm, PSO has to reflect on two important search abilities for achieving high optimal solutions in a short

A. Selamat et al. (Eds.): ACIIDS 2013, Part I, LNAI 7802, pp. 31–40, 2013.

time. These search abilities are diversity and intensity [4]. The former is useful in guaranteeing high accurate results while the latter promises fast processing time.

Original PSO is known with less diversity so that it sometimes converges earlier before achieving high accurate solutions [5]. One promising way for improving PSO diversity is through the insertion of mutation from genetic algorithm (GA) concept [6]. Another technique is by providing time-varying behavior to some of PSO elements or parameters [4]. Based on literature, it is found that time-varying behavior is mostly adapted in determining inertia weight value for PSO. Time-varying is usable to automatically adjust PSO search ability according to iteration number. In this paper, the strengths of both time-varying and mutation are combined in order to attain high synergy of diversity in PSO. Therefore, in order to evaluate the effectiveness of time-varying mutation,the results are compared with time-varying inertia weight.

The remaining content of this paper is organized as follows. In Section 2, brief descriptions on original PSO, time-varying inertia weight and mutation hybridization into PSO are given. Then, the experiments configuration and implementation details are described in Section 3. Section 4 discusses the results before concluding remarks in Section 5.

2 Particle Swarm Optimization

Particle swarm optimization (PSO) is a kind of population based metaheuristics which consists of individual of solutions. The solutions in PSO is represented as particle. These particles are randomly initialized with D-dimensional size that consist a vector of position p, velocity v, personal best fitness (*pbest*) and global best fitness (*gbest*).

Along all iterations, the particles are flying through search space and always accelerated towards better solutions. This process can be achieved by updating the velocity v of each particle i with the following Equation (1):

$$v_{i(t+1)} = v_{i(t)} + c_1 * rand_1 * (pbest - x_{i(t)}) + c_2 * rand_2 * (gbest - x_{i(t)}) \quad (1)$$

where $v_{i(t+1)}$ is velocity of the ith particle at iteration t, $c_1 * rand_1 * (pbest - x_{i(t)})$ is cognitive learning for each particle, $c_2 * rand_2 * (gbest - x_{i(t)})$ is social information of the whole particles, c_1 and c_2 are two positive constants for representing personal and social learning rate respectively and $rand_1$ and $rand_2$ are two separately generated random numbers in the uniform range of [0:1].

Then, the trajectory of position for each particle at each generation $x_{i(t)}$ is updated with the following Equation (2):

$$x_{i(t+1)} = x_{i(t)} + v_{i(t+1)} \quad (2)$$

Since introduction, PSO has gained continuous enhancement from many researchers. Among the variant of PSO, one of the most acceptance kind is inertia weight PSO which has given significant improvement to the performance of PSO. It was introduced by Shi and Eberhart [7] that change the original velocity update in Equation (1) with the following Equation (3):

$$v_{i(t+1)} = wv_{i(t)} + c_1 * rand_1 * (pbest - x_{i(t)}) + c_2 * rand_2 * (gbest - x_{i(t)}) \quad (3)$$

where w is the inertia weight parameter for controlling velocity acceleration of particles i at an iteration t. Shi and Eberhart have claimed that PSO search ability can be dynamically adjusted with different inertia weight value [7]. It has been proved from empirical experiments that the appropriate range for inertia weight is between 0.5 and 0.9. A large value leads the particles to explore wide area search space while a small value encourages particles to deeply explored only on some potential particles.

PSO needs good diversity at the early stages of search which then gradually moves to better intensity at the end of search. With high diversity, particles in PSO are encouraged to explore as many solutions as possible. In contrast, during the end of search, PSO have to concentrate on some potential solutions so that the results for getting optimal solutions can be gained in a short processing time. One promising method to facilitate such dynamic inertia weight value is through applying time-varying strategy.

2.1 Time-Varying Inertia Weight PSO

The time-varying techniques for inertia weight can be either linear or non-linear and increasing or decreasing [4].

Linear decreasing (LD) time-varying have been widely used for inertia weight [7]. In LD technique, the inertia weight value is linearly decreased from its maximum value w_{max} to the smallest w_{min}. Equation (4) defines LD inertia weight:

$$w_t = (iter_{max} - iter)/iter_{max} * (w_{max} - w_{min}) + w_{min} \quad (4)$$

where $iter$ is the current iteration t and $iter_{max}$ is the maximum iterations of PSO search.

In contrast, linear increasing (LI) time-varying has been used in [8]. Experiments have shown that PSO with LI inertia weight technique also perform better than original PSO for some benchmark functions. The formula is:

$$w_t = (iter_{max} - iter)/iter_{max} * (w_{min} - w_{max}) + w_{max} \quad (5)$$

Based on the linear time-varying approaches, some researchers have used non-linear decreasing or increasing. In [9], non-linear decreasing (NLD) inertia weight is defined as the following equation (6):

$$w_t = ((iter_{max} - iter)^n/(iter_{max})^n) * (w_{max} - w_{min}) + w_{min} \quad (6)$$

where n is nonlinear modulation index which the value can be in the range of between $[w_{min} : w_{max}]$. Based on Equation (5), the non-linear increasing (NLI)technique can be defined as:

$$w_t = ((iter_{max} - iter)^n/(iter_{max})^n) * (w_{min} - w_{max}) + w_{max} \quad (7)$$

Another technique for NLI named Sugeno was proposed in [10] with the following Equation (8).

$$w_t = w_{min} * u^{iter} \tag{8}$$

where w_{min} is the initial inertial value selected in the range of [0:1] and u is a constant value in the range of [1.0001:1.005]. A chaotic model time-varying inertia weight is proposed by [11] that incorporating chaotic item into the linear decreasing technique. The formula for chaotic model time-varying is defined as:

$$w_t = (iter_{max} - iter)/iter_{max} * (w_{max} - w_{min}) + w_{min} * z \tag{9}$$

where $z = 4z(1 - z)$ and its initial value is randomly generated within the range of $[0 : 1]$.

Responding to the remarkable performance of time-varying inertia weight, this research try to adapt the time-varying techniques into mutation implementation.

2.2 PSO Hybridization with Mutation

Based on GA concept, mutation is a genetic operator that alters one or more gene values in chromosome of GA. With the new gene values, GA might be able to arrive at better solution than was previously possible. Mutation is an important part of the genetic search as it helps to prevent the population from stagnating at any local optima. Responding to the strength, mutation has been widely hybridized into PSO. As for example, Higashi and Iba [12] have used Gaussian distribution mutation operation for mutation modifying position of some selected particles. The mutation operation is defined as the following Equation (10).

$$mutatex_{i(d)} = x_{i(d)} + Gaussian(\alpha) \tag{10}$$

where $Gaussian(\alpha)$ is a function that returns a random value from Gaussian distribution. The α value is bounded within 0.1 times of the particle dimension size.

Stacey et. al [13] has proposed similar mutation operation in Equation (9) but changes the Gaussian into Cauchy distribution. Therefore, Equation (9) has been altered to be as Equation (11).

$$mutatex_{i(d)} = x_{i(d)} + Cauchy(\alpha) \tag{11}$$

In other works, Michalewiczs non-uniform mutation operator has been used for generating random numbers in relation to iteration number [6]. The mutation changes the dth position of the ith particle at the tth iteration with the following Equation (12).

$$mutatex_{i(d)} = \begin{cases} x_{i(d)} + delta(t, U - x_{i(d)}) \; if \; rb = 1 \\ x_{i(d)} + delta(t, x_{i(d)} - L) \; if \; rb = 0 \end{cases} \tag{12}$$

where U and L is the maximum and minimum position value of the particle dimension. A value 1 or 0 is randomly generated for rb and the *delta(t,y)* is a function defined as:

$$delta(t, y) = y(1 - 1^{(1-\frac{t}{T})^{b}}) \tag{13}$$

where r is a random number, t is the current iteration, T is the total iteration and b is a parameter that has been set to 5.

There is also a technique that has applied triggered method to detect particle conditions either weak or healthy [14]. The particle is considered as weak when there are available particles remaining their fitness at a sequence number of iterations. For example, when the percentage of particles in healthy condition is less than sixty percent for four times of iteration, the following uniform mutation operations into particle trajectory is performed.

$$mutatex_{i(d)} = x_{i(d)} + (p - 0.5) * x_{i(d)} \tag{14}$$

$$mutatev_{i(d)} = v_{i(d)} + (p - 0.5) * v_{i(d)} \tag{15}$$

where $p \in (1, 0)$ is a random value, $x_{i(d)}$ is the position and $v_{i(d)}$is the velocity of dth dimension in the ith particle. The mutation on each particle dimension is implemented only when the generated random number $r \in (1, 0)$ is less then mutation probability R.

Hybridizing mutation into PSO with triggered method is categorized as dynamic strategies that are automatically implemented according to current PSO condition. The results from empirical experiments have revealed that dynamic mutation can significantly increase PSO performance towards achieving high accurate solutions[4]. Recently, it is found that current techniques for dynamic mutation for PSO were mostly implemented with self-adaptive strategies rather than time-varying[4].

Inspired from the remarkable achievement gained by PSO with time-varying inertia weight, this research interest has been directed towards incorporating time-varying mutation into PSO.

3 Experiments

The experiments try to apply different strategies of time-varying inertia weight into mutation implementation. These strategies are named as S1 for linear increasing, S2 for linear decreasing, S3 for Sugeno, S4 for non-linear increasing, S5 for non-linear decreasing and S6 for chaotic. Besides, all experiments are divided into two groups namely time-varying mutation and time-varying inertia weight. The flowchart for each implementation is illustrated in **Fig. 1**.

As shown in **Fig. 1a**, Time-varying mutation is implemented for determining dynamic value of mutation weight in mutation operation. Therefore, different iteration numbers use different mutation weight. The probability for each particle position to be chosen for mutation depends on 0.5 mutation rate. The mutation operation for each selected position from ith particle is defined as:

$$mutatex_{i(d)} = x_{i(d)} + M + Gaussian(\alpha) \tag{16}$$

where $Gaussian(\alpha)$ returns random value within the size of particle dimension, $x_{i(d)}$ is the selected ith particle at position d and M denotes mutation weight. The Gaussian mutation have been widely used in PSO hybridization [12].

Then, the set of experiments for time-varying inertia weight has been implemented without mutation element based on the illustration in **Fig. 1b**. Time-varying inertia weight uses Equation (3) for velocity update and Equation (2)

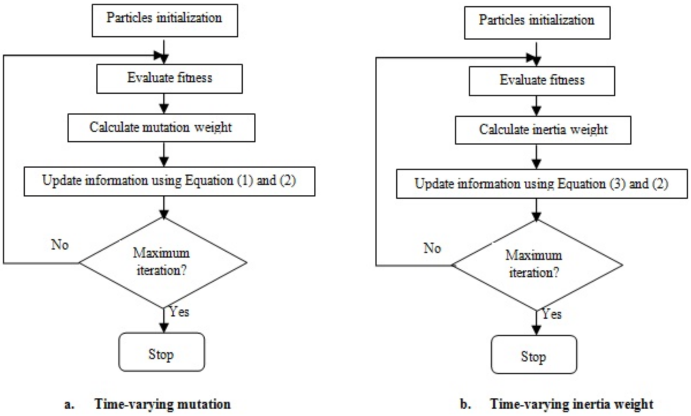


Fig. 1. Time-varying for mutation and inertia weight

for position update. The maximum and minimum value for inertia weight is set to 0.4 to 0.9 respectively.

3.1 General Configuration

Each of experiment is carried out with 2000 iterations from 30 times of independent runs. The number of particles is set to 40 while the size of dimension is bounded with 30. The personal and social learning rate (c_1 and c_2) are set to be 2.0. To illustrate the effectiveness and performance of self-adaptive PSO for optimization problems, a set of four representative benchmark functions are employed. It is encountered that these benchmark functions have been widely used for representing real optimization problems. The functions are listed in Table 1.

Table 1. Benchmark functions

Function	Name	Search dimension	Initialization dimension
$f_1(x) = \sum_{i=1}^{D} ix_4^i + rand$	Noisy	$[-1.28 : 1.28]$	$[0 : 0]$
$f_2(x) = \sum_{i=1}^{D-1}[100(x_i^2 - x_{i+1})^2 + (x_i - 1)^2]$	Rosebrock	$[-5 : 10]$	$[0 : 0]$
$f_3(x) = 10D + \sum_{i=1}^{D-1}[x_i^2 - 10cos(2\pi x_i)]$	Rastrigin	$[-5.12 : 5.12]$	$[1 : 1]$
$f_4(x) = sin^2(\pi y_1) + \sum_{i=1}^{D-1}[(y_i - 1)^2(1 + 10sin^2(\pi y_i + 1))] + (y_D - 1)^2(1 + sin^2(2\pi x_D))$ and $y_i = 1 + \frac{x_i - 1}{4}$	Levy	$[-10 : 10]$	$[1 : 1]$

4 Results

In this section, the performance of each time-varying strategy (S1 to S6) is compared according to the different implementations (mutation and inertia weight). The metrics used in the evaluation are best fitness optimal and convergence speed.

4.1 Best Fitness Optimal

Table 2 shows that the time-varying strategies for mutation have produced high optimal solutions in all of the benchmark functions. In Rastrigin function (f_3), all strategies have achieved good optimal results with similar value of -0.3. S2 strategy is shown to be superior than other strategies in optimizing Levy functions (f_4), S3 is able to generate better results in Rosenbrock fucntion (f_2) and S6 has produce remarkable achievement in Noicy function (f_1).

Table 2. Mean best fitness with time-varying mutation operation

Function	S1	S2	S3	S4	S5	S6
f_1	$6.7E-04$	$7.1E-04$	$5.5E-04$	$5.7E-04$	$5.2E-04$	**4.6E-04**
f_2	$2.6E-02$	$2.5E-02$	**-3.0E-01**	$2.0E-02$	$1.7E-02$	$2.6E-02$
f_3	**-3.0E-01**	**-3.0E-01**	**-3.0E-01**	**-3.0E-01**	**-3.0E-01**	**-3.0E-01**
f_4	$5.1E-02$	**3.7E-02**	$4.7E-02$	$4.2E-02$	$5.1E-02$	$6.8E-02$

Table 3 shows the mean best fitness of all time-varying strategies in inertia weight. The results show that all functions especially Levy (f_4) have achieved less optimal results with most of the time-varying strategies in inertia weight as compared to time-varying mutation. Besides, the optimal results from S1 and S6 strategies in inertia weight are tremendously lower than mutation for all functions. For example, mean best fitness for Noicy function (f_1) is 0.048 with S1 inertia weight but it is 0.00067 with S1 mutation. In addition, the mean best fitness for Rastrigin (f_3) function has been slightly increased by using S1 and S6 strategies from -0.3 in time-varying mutation into -0.2 in time-varying inertia weight. As for better illustration, **Fig. 2** is drawn to represent the different of optimal results between time-varying mutation and time-varying inertia weight.

Table 3. Mean best fitness with time-varying inertia weight

Function	S1	S2	S3	S4	S5	S6
f_1	$4.8E-02$	$3.0E-04$	**7.8E-05**	$4.6E-03$	$1.6E-04$	$3.9E-01$
f_2	$0.0E+00$	**1.2E-04**	$8.9E-04$	$0.0E+00$	$0.0E+00$	$0.0E+00$
f_3	$-2.5E-01$	**-3.0E-01**	**-3.0E-01**	**-3.0E-01**	**-3.0E-01**	$-2.4E-01$
f_4	$4.0E+00$	$1.8E+00$	$2.2E+00$	**5.3E-03**	$2.5E-01$	$1.1E+00$

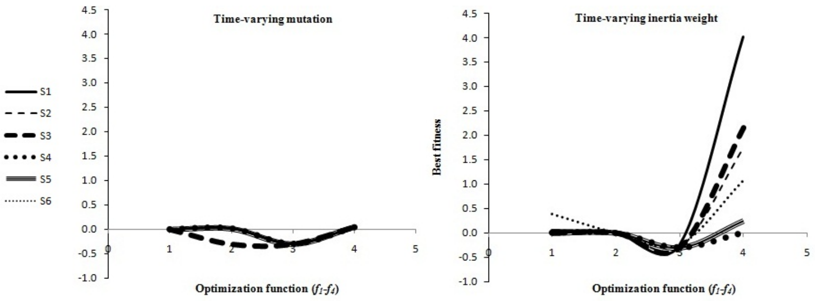


Fig. 2. Mean best fitness of the benchmark functions

As seen in the **Fig. 2**, time-varying mutation is generally performs better in producing high optimal solutions than time-varying inertia weight. In time-varying inertia weight, some strategies have produced more than 1.0 of mean best fitness and S1 strategy has generated the highest value for Levy function (f_4) with 4.0 of mean best fitness. In contrast, the values for all time-varying strategies in mutation are extremely low (less than 0.5). These results indicate that PSO search diversity can be increased through hybridizing PSO with time-varying mutation.

4.2 Convergence Rate

Fig. 3 provide general comparison of convergence rate performance between time-varying mutation and time-varying inertia weight. It can be seen that most of the time-varying strategies in mutation have produced better convergence rate than time-varying inertia weight. This improvement is drastically depicted in the two multi-modal functions (Rastrigin and Levy). In Levy function (f_4),

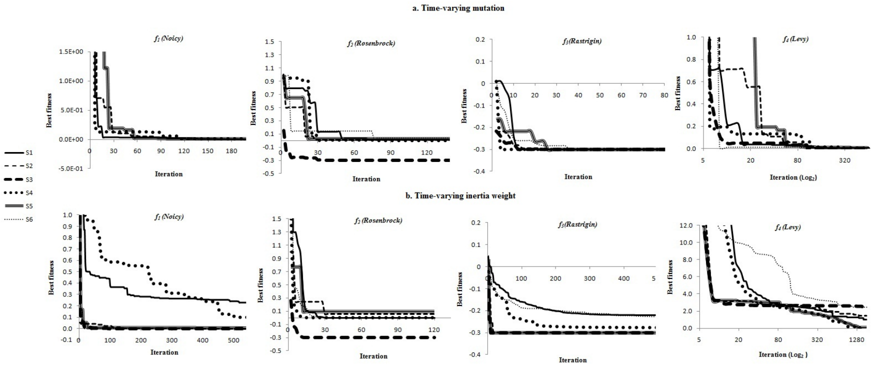


Fig. 3. Mean convergence rate of the benchmark functions with different strategies of time-varying mutation

the convergence rate from all time-varying strategies in mutation are less than 300 iterations which is faster than time-varying strategies in inertia weight (more than 1000 iterations). Besides, an extreme improvement can be achieved in Noicy function (f_3) where the convergence rate for all time-varying strategies in mutation are less than 50 iterations but all time-varying strategies in inertia weight have generated the optimal results only after more than 400 iterations. Some strategies (S1 and S4) in Noicy function (f_1) have received remarkable improvement to the convergence rate when implementing time-varying mutation. The rate is less than 100 iterations with time-varying mutation but more than 500 iterations with time-varying inertia weight.

5 Conclusion

Based on the empirical experiments, it has been discovered that the implementation of time-varying strategies into mutation has generally increased PSO performances in relation to best fitness optimal and convergence rate. The improvement can be significantly shown to the multi-modal functions that consists of many local optima but only one is the global optimum. In optimizing multimodal functions, the search algorithm are facing difficulty to converge from the many local optima. Nevertheless, this research has revealed that time-varying mutation is capable to resolve the problem. On the other hand, this techniques are able to increase both PSO search abilities namely intensity and diversity.

By introducing time-varying mutation, this paper has opened up many possibilities for PSO future improvements. In other research works, these time-varying strategies should be analyzed with the implementation of *selection* and *crossover*.

Acknowledgement. The authors would like to thank Kementerian Pengajian Tinggi MALAYSIA and Universiti Teknologi MARA for their financial support to this project.

References

[1] Kennedy, J., Eberhart, R.: Particle swarm optimization. In: IEEE International Conference on Neural Networks, vol. 4, pp. 1942–1948 (1995)

[2] Bratton, D., Kennedy, J.: Defining a standard for particle swarm optimization. In: IEEE Swarm Intelligence Symposium, pp. 120–127 (April 2007)

[3] Thangaraj, R., Pant, M., Abraham, A., Bouvry, P.: Particle swarm optimization: Hybridization perspectives and experimental illustrations. Applied Mathematics and Computation 217(12), 5208–5226 (2011)

[4] Nickabadi, A., Ebadzadeh, M.M., Safabakhsh, R.: A novel particle swarm optimization algorithm with adaptive inertia weight. Applied Soft Computing (2011)

[5] Yang, X., Yuan, J., Yuan, J., Mao, H.: A modified particle swarm optimizer with dynamic adaptation. Applied Mathematics and Computation 189(2), 1205–1213 (2007)

[6] Andrews, P.S.: An Investigation into Mutation Operators for Particle Swarm Optimization. In: Evolutionary Computation, pp. 1044–1051 (2006)

[7] Eberhart, R., Shi, Y.: Comparing inertia weights and constriction factors. In: IEEE Congress on Evolutionary Computaton, CEC 2000, San Diego, pp. 84–89. IEEE (2000)

[8] Zheng, Y., Ma, L., Zhang, L.: Empirical study of particle swarm optimizer with an increasing inertia weight. In: IEEE Congress on Evolutionary Computation, pp. 221–226 (2003)

[9] Chatterjee, A., Siarry, P.: Nonlinear inertia weight variation for dynamic adaptation in particle swarm optimization. Computers & Operations Research 33(3), 859–871 (2006)

[10] Jiao, B., Lian, Z., Gu, X.: A dynamic inertia weight particle swarm optimizatin algorithm. Chaos, Solitons & Fractals 37, 698–705 (2008)

[11] Feng, Y., Yao, Y., Wang, A.: Comparing with chaotic inertia weights in particle swarm optimization. In: International Conference on Machine Learning and Cybernetics, pp. 329–333 (2007)

[12] Higashi, N., Iba, H.: Particle swarm optimization with Gaussian mutation. In: IEEE Swarm Intelligence Symposium, SIS 2003, pp. 72–79 (2003)

[13] Stacey, A., Jancic, M., Grundy, I.: Particle swarm optimization with mutation. In: The 2003 Congress on Evolutionary Computation, CEC 2003, vol. 2, pp. 1425–1430 (2003)

[14] Zhou, Y., Tan, Y.: Particle swarm optimization with triggered mutation and its implementation based on GPU. In: Proceedings of the 12th Annual Conference on Genetic and Evolutionary Computation, GECCO 2010, pp. 1–8 (2010)

Extending and Formalizing Bayesian Networks by Strong Relevant Logic

Jianzhe Zhao[1,2], Ying Liu[2], and Jingde Cheng[3]

[1] School of Business Administration, Northeastern University, Shenyang 110819, P.R. China
[2] Software College, Northeastern University, Shenyang 110819, P.R. China
{zhaojz,liuy}@swc.neu.edu.cn
[3] Department of Information and Computer Sciences, Saitama University,
Saitama, 338-8570, Japan
cheng@ics.saitama-u.ac.jp

Abstract. In orders to deal with uncertainty by systematical methodologies, some structural models combining probability theory with logic systems have been proposed. However, these models used only the formal language part of the underlying logic system to represent empirical knowledge of target domains, but not asked the logical consequence theory part of the underlying logic system to reason about empirical theorems that are logically implied in domain knowledge. As the first step to establish a unifying framework to support uncertainty reasoning, this paper proposes a new framework that extends and formalizes traditional Bayesian networks by combining Bayesian networks with strong relevant logic. The most intrinsic feature of the framework is that it provides a formal system for representing and reasoning about generalized Bayesian networks, and therefore, within the framework, for given empirical knowledge in a specific target domain, one can reason out those new empirical theorems that are certainly relevant to given empirical knowledge. As a result, using an automated forward reasoning engine based on strong relevant logic, it is possible to get Bayesian networks semi-automatically.

Keywords: Uncertainty reasoning, Bayesian networks, Strong relevant logic, Forward reasoning, Free-EnCal.

1 Introduction

Uncertainty is one of the most intrinsic characteristics of the real world and therefore is investigated in various disciplines. How to deal with uncertainty is still a challenging problem in Artificial Intelligence as well as Computer Science. In order to deal with uncertainty by systematical methodologies based on fundamental theories, some structural models combining probability theory with logic systems have been proposed.

Bayesian networks (BNs) are very important tools for representing and reasoning about uncertainty using probability theory and graph theory. Bayesian networks use directed acyclic graphs (DAGs) to provide an explicit representation for conditional dependences of random variables that are relevant to each other in a given state of

A. Selamat et al. (Eds.): ACIIDS 2013, Part I, LNAI 7802, pp. 41–50, 2013.

domain knowledge [1,2]. As a probabilistic network, it offers an efficient means of representing the probability functions, and drawing inferences from these probability functions [3]. However, BNs do not used formal language to represent empirical knowledge of target domains. In order to represent and reason about uncertainty more formally, we need a new framework that extends traditional Bayesian networks by asking some logic system to represent and reason about empirical theorems that are logically implied in domain knowledge.

Several probabilistic logical models (PLMs) have been proposed by combining BNs with Logic Programming, first-order logic, or relational languages [4]. The most developed models are Bayesian Logical Programs (BLPs) [5] and Logical Bayesian Networks (LBNs) [6]. However, from the viewpoint of application, there is a crucial problem in these models: the underlying logic system (i.e., classical mathematical logic, CML) these models based on is not appropriate for practical applications in the following senses. First, judging from the aspect of representation, the logical languages of the models did not use the full language of the underlying logic system but only some incomplete language facilities with various restrictions. Second, judging from the aspect of reasoning, these models used only the formal language part of the underlying logic system to represent empirical knowledge of target domains, but not asked the logical consequence theory part of the underlying logic system to reason about empirical theorems that are logically implied in domain knowledge. In facts, as many investigations about roles of logic in various applications showed in literature, these problems come from classical mathematical logic and its limitations [7-12].

As the first step to establish a unifying framework to support uncertainty reasoning, this paper proposes a new framework that extends and formalizes traditional Bayesian networks by combining Bayesian networks with strong relevant logic [7,8]. The most intrinsic feature of the new framework is that it provides a formal system based on strong relevant logic for representing and reasoning about generalized Bayesian networks, and therefore, within the framework, for given empirical knowledge in a specific target domain, one can reason out those new empirical theorems that are certainly relevant to given empirical knowledge. As a result, using an automated forward reasoning engine based on strong relevant logic [13], it is possible to get Bayesian networks semi-automatically.

The rest of this paper is organized as follows: Section 2 presents basic notions and terminologies of Bayesian networks and strong relevant logic, Section 3 formally defines strong relevant logical Bayesian networks, Section 4 discuss related works briefly, and concluding remarks are given in Section 5.

2 BNs and SRL: Basic Notions and Terminologies

2.1 Bayesian Networks

A BN consists of two components, i.e., qualitative and quantitative component.

Qualitative component contain nodes in a DAG, i.e. random variables and directed edges. We use X_i to denote a node; it is an element in a finite set $X = \{X_1, \dots, X_n\}$, which corresponds to a set of random variables. A random variable can take continuum or discrete values; we believe that in modeling actual problems discrete random variables are special cases. Directed edges connected nodes; if there is a directed edge from X_i to X_j $(i, j \leq n; i \neq j)$, then we call X_i is a parent of X_j, and X_j is a descendant of X_i, the parents of X_i are denoted by $\pi(X_i)$. If X_i has no parents, it is a root node.

Quantitative component is parameters quantifying influence between $\pi(X_i)$ and X_i. The parameters of a BN are conditional probability distributions (CPDs) of every node. Probability distributions of root nodes are $P(X_i)$, others are $cpd(X_i) = P(X_i \mid \pi(X_i))$. A BN represents the knowledge of the problem domain as a joint probability distribution:

$$P(X_1,\ldots,X_n) = \prod_{i=1}^{n} P(X_i|\pi(X_i))\cdot$$

Bayesian networks have a basic assumption: there must have a subset of nodes $A \subseteq X$ for arbitrary X_i, which are not descendant nodes of X_i; for any given joint states of $\pi(X_i)$, X_i is conditionally independent of A, i.e. $P(X_i \mid A, \pi(X_i)) = P(X_i \mid \pi(X_i))$.

2.2 Strong Relevant Logic

Traditional relevant logics ware constructed during the 1950s-80s in order to find a mathematically satisfactory way of grasping the elusive notion of relevance of antecedent to consequent in conditionals, and to obtain a notion of implication (i.e., relevant implication) which is free from the so-called "paradoxes" of material implication in classical mathematical logic and "paradoxes" of strict implication in Lewis's modal logics [9-12]. One of the most intrinsic features of the relevant logics is that they have a primitive intensional connective (relevant implication or entailment) to represent the notion of conditional and their logical theorems include no implicational paradoxes. The underlying principle of the relevant logics is the following relevance principle:

The relevance principle: If A⇒B, where ⇒ denotes the notion of relevant implication or entailment, is a logical theorem of a relevant logic, for any two propositional formulas A and B, then A and B share at least one propositional variable.

Strong relevant logics ware constructed in order to establish a satisfactory logic calculus of conditional to underlie relevant reasoning [7,8]. The logics require that the premises of an argument represented by a conditional include no unnecessary and needless conjuncts and the conclusion of that argument includes no unnecessary and needless disjuncts. As a modification of traditional relevant logics, strong relevant logics reject conjunction-implicational paradoxes and disjunction-implicational paradoxes in traditional relevant logics. What underlies the strong relevant logics is the following strong relevance principle:

The strong relevance principle: If A is a logical theorem of a strong relevant logic, then every sentential variable in A occurs at least once as an antecedent part and at least once as a consequent part in A.

Today, relevant logic has become an important branch of philosophical logic. It is the only family of logics to deal with the relevant account of validity in reasoning. As a knowledge representation and reasoning tool, relevant logic has many useful properties that the classical mathematical logic and its various classical conservative extensions do not have, and therefore, it is most hopeful candidate as the fundamental logic to underlie those advanced information and/or knowledge systems where the relevant reasoning plays a crucial role.

3 Extending Bayesian Networks by Strong Relevant Logic

In this section we introduce Strong Relevant Logical Bayesian Networks (SRL-BNs) that are formal systems combing BNs and SRL. We discuss the motivation in the Section 3.1 and then give its definition and declarative semantics in Section 3.2. Consider the following running example, similar to the university example in Probabilistic Relational Models [18].

There are students, teachers and courses. Each student has an IQ, each teacher has a level and each course is either interesting or not. A student selecting a certain course gets a grade for that course. The interestingness of a course is influenced by the sum of the IQ of all students selecting the course and the level of the teacher opening the course. A student's grade for a course is influenced by the student's IQ. The specific knowledge may contain which students select which courses, which teacher open which course, level of each teacher, IQ of each student.

Suppose that the domains of random variables are courses { *ma, ch, mu* }, students { *li, liu, wang* }, teachers { *pete, marry* }, iq { *0,...,120* }, level { *1, 2*}, interesting { *0,1* } and grade { *a, b, c, d* }. If the specific situations of the problem contain: *li* selects *ma* and *ch*, *liu* selects *ma*, *wang* selects *mu*; and *pete* opens *ma* and *ch*, *marry* opens *mu*. Based on the knowledge we can determine a BN of the specific situations with the structure below.

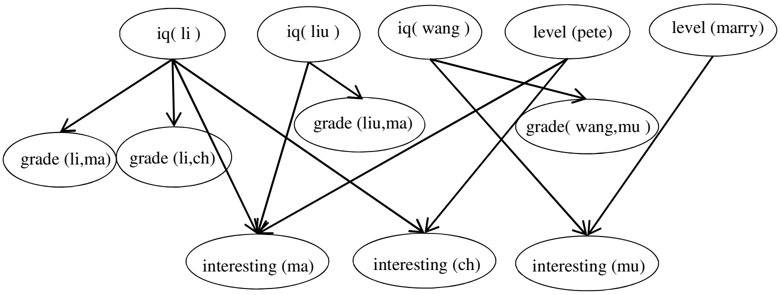


Fig. 1. A Bayesian network of running example

3.1 Motivation of Introducing SRL-BNs

We compare SRL-BNs with BLNs and LBNs by our running example.

The advantage of BLPs is that they use formal language part of the underlying logic system to represent generalized BNs. They separate the quantitative and qualitative aspects of the model successfully; therefore they are not rigid structures anymore. In BLPs, the Bayesian clauses are used to describe BNs, each of which contains part of head and body. However, there are also some disadvantages in modeling real-word problems. First, it is very difficult to not only generate the Bayesian networks they want but also interpret the clauses [5]. Second, Bayesian clauses need some restrictions to avoid derivation of non-ground true facts.

In BLPs, predicates are called Bayesian predicates and they are different from ordinary logical predicates because they have associated ranges [5]. If we model running example with a BLP, the Bayesian predicates are student/1 {true, false}, teacher/1 {true, false}, course/1 {true, false}, iq/1 {0,…,120}, level/1 {1,2}, select/2 {true, false}, open/2 {true, false}, interesting/1 {0,1} and grade/2 {a,b,c,d}. Some ground Bayesian atoms like student(li), teacher(pete), select(li,ma), etc., are not nodes in the DAG of the running example(See Fig 1). We need the Bayesian clauses below to represent our general information.

iq(s) | student(s) .
level(t) | teacher(t) .
grade(s,c) | iq(s), select(s,c) .
interesting(c) | iq(s), select(s,c), level(t), open(t,c) .
interesting(c) | course(c) .

For an instance of our example, if interesting(ch) is denoted by X_i, iq(li), select(li,ch), level(pete), open(pete,ch) and course(ch) are $\pi(X_i)$, but it is inconsistent with the DAG we obtain.

LBNs provide a more intuitive way of modeling real-word problems, but the general knowledge which determines a Bayesian network contains empirical knowledge of target domains. As a consequent, a LBN is a formal language for a specific target domain. Another disadvantage is that human plays the most important role in modeling. While learning form literals, humans avoid derivation of non-ground true facts by themselves but not by logical consequence theory part.

Predicates of LBNs contain logical literals and probabilistic predicates. The logical literals are ordinary logical predicates and the probabilistic predicates are similar to Bayesian predicates in BLPs [6]. If we model running example with a LBP, the clauses of a LBN for the running example below, where student/1, course/1, select/2, open/2 and teacher/1 are logical literals; and iq/1 {0,…,120}, level/1 {1,2}, grade/2 {a,b,c,d} and interesting/1 {0,1} are probabilistic predicates. Human decide either logical literals or probabilistic predicates from the context of the literals, and the process is relevant to specific domain.

The clauses of LBNs contain random variable declarations and conditional dependency clauses [6]. The random variable declarations for running example are:

iq(s) ← student(s) .
interesting(c) ← course(c) .
grade(s,c) ← select(s,c) .
level(t) ← teacher(t) .

And the conditional dependency clauses are:

interesting(c) | iq(s), level(t) ← select(s,c), open(t,c) .
grade(s,c) | iq(s) .

For the first conditional dependency clause, it is difficult to interpret, and the relevance of clause is considered by human cognition but not logic.

3.2 Defining SRL-BNs

First we define SRL-BNs and its components, give its declarative semantics, and then model the running example by them.

Definition 1 (Qualitative Bayesian Clauses). A Qualitative Bayesian Clause (Qual-BC) is an expression of the form

$$A \mid C_1 \wedge \cdots \wedge C_n .$$

where $n \geq 0$, C_i is simply conditional with an expression of the form $B_i \Rightarrow A_i$ ($1 \leq i \leq n$), A and A_i are Bayesian atoms, B_i is conditional atoms, A_i and B_i are called the consequent and the antecedent of C_i; a set of logical predicates containing in Qual-BCs decide Bayesian atoms and conditional atoms.

Definition 2 (Procedure Bayesian Clause). A Procedure Bayesian Clause is a clause of the form

$$A \mid A_1, \ldots, A_n .$$

Where A_i ($1 \leq i \leq n$) is the consequent of C_i, A and $A_1, \ldots, A_n$ ($n \geq 0$) are Bayesian atoms and all atoms are universally quantified.

Definition 3 (Quantitative Bayesian Clauses). A Quantitative Bayesian Clause (Quan-BC) c is an expression of the form

$$r(t_1, \ldots, t_n) \mid s_1(t_{1,1}, \ldots t_{1,n_1}), \ldots, s_m(t_{m,1}, \ldots t_{m,n_m}).$$

A Quantitative Bayesian Clause corresponds to a Procedure Bayesian Clause and with the conditional probability distribution of c denoting as $cpd(c)$. For atoms $r(t_1, \ldots, t_n)$, the domain $D(r)$ is unique for the probabilistic predicate r. The $cpd(c)$ specify for each $(u_i, \ldots, u_m) \in D(s_1) \times \ldots \times D(s_m)$ a function with

$$cpd(c)(u \mid u_1, \ldots, u_m) = p(r(t_1, \ldots, t_n) = u \mid s_1(t_{1,1}, \ldots t_{1,n_1}) = u_1, \ldots, s_m(t_{m,1}, \ldots t_{m,n_m}) = u_m).$$

Definition 4 (Combined Conditional Probability Distributions, Comb-CPDs). A Comb-CPD for a probabilistic predicate r is any algorithm mapping a set of ground Bayesian atoms to a conditional probability distribution on $D(r)$.

Definition 5 (Strong Relevant Logical Bayesian Networks, SRL-BNs). A SRL-BN is a finite set of Qual-BCs and Quan-BCs. Qual-BCs contain a set of logical predicates and decide qualitative components of Bayesian networks; Quan-BCs are Procedure Bayesian Clauses with CPDs; contain a set of probabilistic predicates; and for each probabilistic predicate it associates a Comb-CPD. Quan-BCs decide quantitative components of Bayesian networks.

The declarative semantics of SRL-BNs is that it defines formal systems extending and formalizing Bayesian Networks by SRL-BNs.

Definition 6 (Extending and Formalizing Bayesian Networks). A BN extending and formalizing by a SRL-BN consist of quantitative and qualitative components.

Where qualitative components (DAG) contain:

- A node, i.e. random variable X_i iff X_i is a ground Bayesian atom in Qual-BCs.
- A directed edge from a node of $\pi(X_i)$ denoted by X_π to X_i in the DAG iff there is a ground instance $X_i | B_1 \Rightarrow A_1, \ldots, B_n \Rightarrow A_n$. of a clause in Qual-BCs such that $X_\pi \in \{A_1, \ldots, A_n\}$and the Qual-BCs are true in the specific SRL-BN.

And quantitative components are CPDs of a DAG:

By applying the Comb-CPD for X_i in Quan-BCs, the CPD for a node X_i is obtained to the set of ground probabilistic atoms that are $\pi(X_i)$ in the DAG.

Definition 7 (Independent Assumption). Each node X_i in a SRL-BN, a subset of nodes $A \subseteq X$ in the SRL-BN, which are not descendant nodes of X_i; for any given joint states of $P(A)$, X_i is conditionally independent of A, i.e. $P(X_i | A, P(A)) = P(A | P(A))$.

Definition 8 (Well-defined SRL-BNs). A SRL-BN is well-defined if and only if it is not empty; the DAG is acyclic; and the ancestor of each node has a finite number.

For running example, the simply conditional are select(s,c) ⇒ iq(s) and open(t,c) ⇒ level(t); the logical predicates are interesting, iq/1, select/2, level/1, open/2 and grade/2; and the probabilistic predicates are iq/1 {0,…,120}, level/1 {1,2}, grade/2 {a,b,c,d} and interesting/1 {0,1}; there are two Qual-BCs like:

interesting(c) | select(s,c) ⇒ iq(s) ∧ open(t,c) ⇒ level(t) . grade(s,c) | iq(s) .

The first clause states that "whether a course is interesting depends on the IQ of a student, if the student selects the course; and depends on the level of a teacher, if the teacher opens the course". In the clause, select(s,c) ⇒ iq(s) and open(t,c) ⇒ level(t) are simply conditionals, where iq(s) and level(t) are the consequents, select(s,c) and open(t,c) are the antecedents. The Qual-BCs decide two Procedure Bayesian Clauses:

interesting(c) | iq(s), level(t) . grade(s,c) | iq(s) .

Each Procedure Bayesian Clause corresponds to a Quan-BC. An assumption for our example, Quan-BC *c* is interesting(c) | iq(s), level(t) .with the *cpd*(*c*) in the table 1.

Table 1. An assumption of the *cpd(c)*

iq(s)	level(t)	interesting(c)=1	interesting(c)=0
0<sum iq(s)<1000	1	0.5	0.5
0<sum iq(s)<1000	2	0.1	0.9
sum iq (s)≥1000	1	0.9	0.1
sum iq (s)≥1000	2	0.5	0.5

A substitution of *c* specifies the conditional probability distribution in the running example: *p*(interesting(ma) l iq(li), iq(liu)¸ level(pete)).For instance, a possible CPD is the following function: If iq(li) + iq(liu)⩾1000 and level(pete) =1 then 0.9/0.1.

By using SRL-BNs, we successfully modeling running example and solve the problems of the way modeling by LBNs and BLPs.

4 Related Works

AI researchers tackling uncertainty can be classified in to three formal schools, which are called logicist, neo-caulist, and neo-probabilist [1]. Probabilistic logics are methods combining aspects of probability theory and aspects of logic, like independent Choice Logic [14], PRISM [15] and Stochastic Logic Programs [16]. Probabilistic networks are kind of the probabilistic logic, and Bayesian networks are simply probabilistic networks [3], noting that there are possibilistic networks [17]. For solving some fundamental problems of BNs, researchers combine Bayesian networks with logics. The research of the paper based on the kind of models. The most developed models are PRMs [18], BLNs and LBNs.

SRL-BNs are related to BLPs as expressive formal systems representing Bayesian networks, they do not have a rigid structure any more, i.e. are not used to generate specific Bayesian networks for particular queries, they separate the quantitative and qualitative aspects of the model successfully.

SRL-BNs are related to LBNs, because the frameworks of LBNs are intuitive and they are always easy to interpret. They are useful language for modeling specific problems.

5 Concluding Remarks

We have proposed a new framework that extends and formalizes traditional Bayesian networks by strong relevant logic. This work is just the first step to establish a unifying framework to support uncertainty reasoning. There are many interesting and challenging research problems and many technical issues for applying the framework to the real world. First, as a theoretical work to investigate formal properties of strong relevant logical Bayesian networks, we should find a satisfactory formal semantics for it. Second, because Bayesian networks can be applied to various areas where modal notions may play important roles, some extensions of strong relevant logical Bayesian networks should be investigated by asking various modal conservative extensions of strong relevant logic [8]. Third, to provide users with easy means to use facilities of the framework, in addition to FreeEnCal, we should design and develop some automated tools to help users to use the facilities easily. Finally, to show the effectiveness and usefulness of the framework in practices, we should perform many case studies to apply the framework to various applications in the real world.

Acknowledgement. Research supported in part by NSFC (61202085), in part by Ph.D. Start-up Foundation of Liaoning Province, China (20111001, 20121002), in part by Fundamental Research Funds for the Central Universities (N110417004).

References

1. Pearl, J.: Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference. Morgan Kaufmann (1988)
2. Pearl, J.: Causality: Models, Reasoning, and Inference, 2nd edn. Cambridge University Press (2009)
3. Haenni, R., Romeijn, J.W., Wheeler, G., Williamson, J.: Probabilistic Logics and Probabilistic Networks. Springer (2010)
4. Fierens, D., Blockeel, H., Bruynooghe, M., Ramon, J.: Logical Bayesian Networks and Their Relation to other Probabilistic Logical Models. In: Kramer, S., Pfahringer, B. (eds.) ILP 2005. LNCS (LNAI), vol. 3625, pp. 121–135. Springer, Heidelberg (2005)
5. Kersting, K., Raedt, L.D.: Bayesian Logic Programs. Technical Report 151, Institute for Computer Science, University of Freiburg, Germany (2001)
6. Fierens, D., Blockeel, H., Bruynooghe, M., Ramon, J.: Logical Bayesian Networks. In: Proceedings of the 3rd Workshop on Multi-Relational Data Mining (MRDM 2004), pp. 19–30 (2004)
7. Cheng, J.: A Strong Relevant Logic Model of Epistemic Processes in Scientific Discovery. In: Information Modeling and Knowledge Bases XI. Frontiers in Artificial Intelligence and Applications, vol. 61, pp. 136–159. IOS Press (2000)
8. Cheng, J.: Strong Relevant Logic as the Universal Basis of Various Applied Logics for Knowledge Representation and Reasoning. In: Information Modeling and Knowledge Bases XVII. Frontiers in Artificial Intelligence and Applications, vol. 136, pp. 310–320. IOS Press (2006)
9. Anderson, A.R., Belnap Jr., N.D.: Entailment: The Logic of Relevance and Necessity, vol. I. Princeton University Press (1975)
10. Anderson, A.R., Belnap Jr., N.D., Dunn, J.M.: Entailment: The Logic of Relevance and Necessity, vol. II. Princeton University Press (1992)
11. Dunn, J.M., Restall, G.: Relevance Logic. In: Gabbay, D., Guenthner, F. (eds.) Handbook of Philosophical Logic, 2nd edn., vol. 6, pp. 1–128. Kluwer Academic (2002)
12. Mares, E.D.: Relevant Logic: A Philosophical Interpretation. Cambridge University Press (2004)
13. Cheng, J., Nara, S., Goto, Y.: FreeEnCal: A Forward Reasoning Engine with General-Purpose. In: Apolloni, B., Howlett, R.J., Jain, L. (eds.) KES 2007, Part II. LNCS (LNAI), vol. 4693, pp. 444–452. Springer, Heidelberg (2007)
14. Poole, D.: The Independent Choice Logic for modeling multiple agents under uncertainty. Artificial Intelligence 94(1-2), 5–56 (1997)

15. Sato, T., Kameya, Y.: PRISM: A symbolic-statistical modeling language. In: Proceeding of the 15th International Joint Conference on Artificial Intelligence (IJCAI 1997), pp. 1330–1335 (1997)
16. Cussens, J.: Parameter estimation in stochastic logic programs. Machine Learning 44(3), 245–271 (2001)
17. Borgelt, C., Gebhardt, J., Kruse, R.: Graphical models. In: Proceedings of International School for the Synthesis of Expert Knowledge, Citeseer (2002)
18. Friedman, N., Getoor, L., Koller, D., Pfeffer, A.: Learning probabilistic relational models. In: Proceedings of the 16th International Joint Conference on Artificial Intelligence (IJCAI 1999), pp. 1300–1309 (1999)

Mining Multidimensional Frequent Patterns from Relational Database

Yue-Shi Lee and Show-Jane Yen

Department of Computer Science & Information Engineering, Ming Chuan University
5 De Ming Rd., Gui Shan District, Taoyuan County 333, Taiwan
{leeys,sjyen}@mail.mcu.edu.tw

Abstract. Mining frequent patterns focus on discover the set of items which were frequently purchased together, which is an important data mining task and has broad applications. However, traditional frequent pattern mining does not consider the characteristics of the customers, such that the frequent patterns for some specific customer groups cannot be found. Multidimensional frequent pattern mining can find the frequent patterns according to the characteristics of the customer. Therefore, we can promote or recommend the products to a customer according to the characteristics of the customer. However, the characteristics of the customers may be the continuous data, but frequent pattern mining only can process categorical data. This paper proposes an efficient approach for mining multidimensional frequent pattern, which combines the clustering algorithm to automatically discretize numerical-type attributes without experts.

Keywords: Data Mining, Clustering, Discretization, Multidimensional Frequent Pattern.

1 Introduction

Modern enterprises and organizations have amassed a lot of data such as market, customer, future trend, etc. and operate in an increasingly competitive environment, such that it is necessary to find and understand customer's behaviors. We can structure massive amounts of data and make data to become useful information. The useful information can help enterprises to produce competitive force. Data mining has been defined as "The nontrivial exaction of implicit, previously unknown, and potentially useful information from data" [4], which can be applied in difference fields. More and more organizations use data mining techniques to find customer knowledge for their marketing strategies.

However, data mining has many different techniques and frequent pattern mining is a popular technique. The discovery of interesting association relationship among business transaction records can help cross-marketing, target-marketing and other business decision. The frequent patterns provide a simple and useful pattern to help companies understand customer behavior and develop selling strategies. Many algorithms about mining frequent patterns have been proposed such as Apriori algorithm [1], FP-growth algorithm [5], etc. However, most researches for mining

A. Selamat et al. (Eds.): ACIIDS 2013, Part I, LNAI 7802, pp. 51–60, 2013.

frequent patterns focus on discovering single dimensional frequent patterns. A typical example of frequent pattern like ("cereal", "milk") would be interpreted as "cereal and milk are purchased frequently" However, single dimensional frequent patterns only provide correlations of the purchased items in a large number of transaction data. This type of rule cannot describe relationships such as "People's age between 20 and 30 with salary between $20000 and $30000 are likely to buy a notebook computer." Enterprises may want to know more information such as customer behavior and customer structure to help them manage customer relationship.

For traditional frequent pattern mining, every record in a transaction database only contains the purchased items. Frequent patterns involve two or more dimensions or predicates can be referred to as multidimensional frequent patterns [6]. Multidimensional association rule can be classified as inter-dimensional association rules and hybrid-dimensional association rules. Inter-dimensional association rules mean that no repeated predicates (dimensions) in the same rule. An example of inter-dimensional association rule like Age (x, "20~30") ∧ Salary (x, "20000-30000") => Buys (x, "notebook computer"). This rule contains three dimensions Age, Salary, Buys. Every dimension only appears once and no repeated predicates are called inter-dimensional association rules. A hybrid association rule contains repeated attributes. An example of hybrid-dimensional association rule is Age (x, "20~30") ∧ Salary (x, "20000-30000") ∧ Buys (x, "notebook computer") => Buys (x, "notebook backpack").

Previous researches proposed methods for mining multidimensional frequent patterns, which transform relational database into a transaction database. These methods always were based on Apriori algorithm [1] and generated a lot of candidates to find frequent itemsets. It needs to scan the database many times and often cost much space. Other method combined classification to discover multidimensional frequent patterns from relational database [7]. Another novel algorithm used many indexes to store multidimensional itemsets and designed a particular structure to discover multidimensional frequent patterns [10], only needs to scan database once and produces index structure to find frequent multidimensional itemsets. But the structure needs much space to store multidimensional itemsets and the designer must have background knowledge to discretize numerical-type attributes.

In this paper, we propose an efficient approach to discover multidimensional frequent patterns. Our approach does not need to scan the original database many times and can progressively reduce the execution time and memory usage. Besides, our approach can also automatically discretize numerical-type attributes without experts. This paper is organized as follows. In section 2, we introduce some related work. The definitions about mining multidimensional frequent patterns are presented in Section 3. Section 4 describes our approach for mining multidimensional frequent patterns, and then we discuss and give a conclusion in section 5.

2 Related Work

There are two types of frequent patterns: single-dimensional frequent pattern and multidimensional frequent pattern. Mining multidimensional frequent patterns involve not only purchased items but also the other attributes such as age, salary and other customer characteristics, etc. In the following, we introduction the two types of frequent patterns.

2.1 Single-Dimensional Frequent Pattern

Various algorithms [1, 3, 4, 8, 11, 12] have been proposed to generate frequent itemsets from a large amount of transaction data. These algorithms generate candidate *k*-itemsets for frequent *k*-itemsets, scan each transaction in a database to count the supports for these candidate *k*-itemsets and find all the frequent *k*-itemsets in the *k*th iteration based on a minimum support threshold. However, because the size of the database can be very large, it is very costly to repeatedly scan the database to count supports for the candidate itemsets. Although FP-Growth algorithm [5] does not need to generate candidate itemsets, it has to take a lot of times to recursively construct many sub-trees which may not fit in main memory when the number of frequent itemsets is large. In order to avoid recursively constructing many sub-trees, COFI-tree algorithm [3] only builds a sub-tree for each frequent item, and generates candidate itemsets and counts their supports from the sub-tree. Since COFI-tree algorithm generates candidate itemsets with any length containing a specific item, the search space for counting the large number of candidates is very large.

2.2 Multidimensional Frequent Pattern

Previous researches transformed a relational database into a transaction database and used Apriori algorithm to mine the multidimensional frequent patterns. Other novel algorithm MDIM (Multi_Dimensional_Indexing_Mining) [10] was not based on Apriori algorithm to discover multidimensional frequent patterns. This algorithm designed a structure that includes indexes to save dimensions and values information. Another method for mining multidimensional frequent pattern was proposed in [9], this method utilized transaction database and customer database to find interesting frequent patterns. They first discovered all the frequent itemsets from the transaction database and recorded customer ID for each frequent itemsets. And then they combine the discovered information with the customer database with conditional attributes. Finally, a relationship graph is used to discover interesting frequent patterns.

Unlike the above method, Chiang and Wu [2] proposed a method that combined a concept hierarchy for each dimension and then produced multidimensional patterns from the concept hierarchical trees. They first divide the database into several element segmentations according to the hierarchy and use Apriori-like algorithm to discover frequent itemsets in every element segmentation, and then combine the segmentations according to the frequent itemsets. Han and et al. Integrate both constraint-based and multidimensional mining into one framework to provide an interactive and exploratory environment for effective and efficient data analysis and mining. They put knowledge constraints, data constraints, dimension constraints, rule constraints or interestingness constraints into mining process to eliminate irrelevant itemsets earlier and minimize the number of the itemsets to be examined.

Lee and Yen combined multidimensional frequent pattern mining and classification. This method proposed a CAM (Classification based on Association Mining) model and discussed attribute dependencies. Because mining frequent patterns cannot process the numerical-type attributes, they used decision tree to discretize the numerical-type attribute values and transformed each numerical-type attribute into categorical type. Finally, the multidimensional frequent patterns are discovered based on Apriori algorithm.

3 Preliminaries

This section introduces some definitions about mining frequent patterns and multidimensional frequent patterns.

3.1 Frequent Pattern

The definitions about frequent patterns are described as follows. A *transaction database* consists of a set of transactions (e.g., Table 1). A *transaction* is a set of items purchased by a customer at the same time. A transaction *t* contains an itemset *X* if every item in *X* is in *t*. The *support* for an itemset is defined as the ratio of the total number of transactions which contain this itemset to the total number of transactions in the database. The *support count* for an itemset is the total number of transactions which contain the itemset. A *frequent pattern* or a *frequent itemset* is an itemset whose support is no less than a certain user-specified minimum support threshold. An itemset of length *k* is called a *k*-itemset and a frequent itemset of length *k* a *frequent k-itemset.*

Table 1. Transaction Database

Transaction ID	Items Purchased
1	Milk, Orange juice, Bread
2	Milk, diapers
3	Soda, diapers, Beer
4	Orange juice, diapers, Beer

3.2 Multidimensional Frequent Patterns

The relational database or data warehouse usually store other attributes not only the items purchased. We can get multidimensional database with m records and n dimensions from the relational database such as Table 2, in which an attribute is regarded as a dimension. There are two types of the attributes in a multidimensional database: numerical attribute and categorical attribute. In Table 2, the five attributes are Occupation, Sex, Target, Age, and Salary. The three attributes Occupation, Sex, and Target are categorical attribute and the two attributes Age, Salary are numerical attributes. In traditional, the original multidimensional database must be transformed into the data which only contains categorical attributes before discovering frequent patterns. A problem is how to discretize the numerical-type attributes and transform each numerical attribute into categorical type, which usually needs background knowledge to determine the range of numerical attribute values.

For a multidimensional database with m records and n dimensions, the dimensions are also called ttributes or fields. All the dimensions can be denoted as $(d_1, d_2, \ldots, d_n)$ in which d_i represents the i-th dimension of the multidimensional database. The j-th record in a multidimensional database can be expressed as $\mathrm{V}ij = (v_{1j}, v_{2j}, \ldots, v_{nj})$ where v_{ij} represents the i-th value in the j-th record, $1 \leqq i \leqq n$ and $1 \leqq j \leqq m$. Therefore, $\mathrm{I}ij = (d_i, v_{ij})$ can be regarded as an item in the multidimensional database. For example, the item (3,1) means that the value in the third dimension (field) and the value is 1. In order to conveniently mine the multidimensional frequent itemsets from the multidimensional database, each categorical attribute value in the multidimensional database can be encoded and the original multidimensional database can be transformed into an encoded multidimensional database. For example, Table 2 can be transformed into Table 3.

Table 2. Original multidimensional database

T_ID	Occupation	Sex	Age	Salary	Paper	Printer
1	Businessman	1	34	38000	1	1
2	Sales Clerk	2	33	36000	1	0
3	Student	1	22	21000	0	1
4	Student	2	23	22000	1	0
5	Businessman	1	34	37000	1	1
6	Sales Clerk	1	24	27000	1	1
7	Sales Clerk	1	30	28000	0	1
8	Businessman	1	28	35000	1	1
9	Student	2	24	25000	1	0
10	Sales Clerk	2	27	32000	1	1

4 Our Approach

In this section, we propose a new method for mining multidimensional frequent patterns. Our method constructs the projected conditional databases for the frequent patterns with length k (k ≥ 1) and use the projected conditional databases to recursively find the multidimensional frequent patterns with length k+1. All the frequent patterns containing a frequent pattern X can be found by growing the frequent pattern X.

Table 3. Encoded multidimensional database

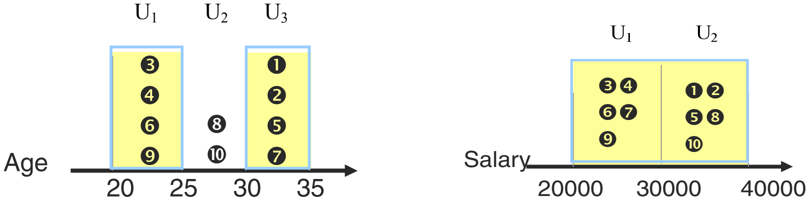
T_ID	Occupation	Sex	Age	Salary	Paper	Printer
1	2	1	34	38000	1	1
2	3	2	33	36000	1	0
3	1	1	22	21000	0	1
4	1	2	23	22000	1	0
5	2	1	34	37000	1	1
6	3	1	24	27000	1	1
7	3	1	30	28000	0	1
8	2	1	28	35000	1	1
9	1	2	24	25000	1	0
10	2	2	27	32000	1	1

Table 4. A transaction database transformed from Table 3

T_ID	Itemset
1	(1,2)(2,1)(3,34)(4,38000)(5,1)(6,1)
2	(1,3)(2,2)(3,33)(4,36000)(5,1)(6,0)
3	(1,1)(2,1)(3,22)(4,21000)(5,0)(6,1)
4	(1,1)(2,2)(3,23)(4,22000)(5,1)(6,0)
5	(1,2)(2,1)(3,34)(4,37000)(5,1)(6,1)
6	(1,3)(2,1)(3,24)(4,27000)(5,1)(6,1)
7	(1,3)(2,1)(3,30)(4,28000)(5,0)(6,1)
8	(1,2)(2,1)(3,28)(4,35000)(5,1)(6,1)
9	(1,1)(2,2)(3,24)(4,25000)(5,1)(6,0)
10	(1,2)(2,2)(3,27)(4,32000)(5,1)(6,1)

The multidimensional database includes categorical attributes and numerical attributes. Because frequent pattern mining can only process categorical attributes, we need to discretize numerical attributes into categorical attributes. Our method applies the clustering algorithm to discretize numerical attributes without experts. We first encode and transform the original multidimensional database (MD) into a transaction database (TD), in which each record in MD is transformed into an (attribute, value) pair which can be regarded as an item and a set of (attribute, value) pairs is an itemset. For example, Table 2 is an original multidimensional database, and Table 3 is an encoded multidimensional database in which each categorical attribute value is encoded. The transformed transaction database is shown in Table 4 in which (attribute, value) pair (i, x) represents that the value of the i-th attribute is x.

Our algorithm automatically discretizes numerical attribute values as follows. According to user-specified *segment size*, we can segment the space formed by the numerical attribute into many units. For the above example, suppose the segment size is set to be 5 for the attribute "Age". Each unit contains five years and there are three units $U_1 = \{[20, 25)\}$, $U_2 = \{[25, 30)\}$, and $U_3 = \{[30, 35)\}$ for this example. Suppose the segment size is 10000 for attribute “Salary”. There are two units U_1 = {[20000, 30000)}, U_2 = {[30000, 40000)} in this example. After segmenting the space into units, the *density threshold* needs to be specified to find *dense units*. If a unit is a dense unit, then this unit may become a portion of a cluster. For example, in Figure 1, the gray units are dense units if the density threshold is set to be 30%, that is, there at least are 30%×10=3 records in the unit. A unit is not high dense unit because density is not greater than density threshold.



(a) The dense units for Age

(b) The dense units for Salary

Fig. 1. The dense units for the numerical attribute values in Table 4

We first choose the starting value in the first unit, and then expand the values in the current unit and the adjacent dense units, until a non-dense unit is reached. A range can be formed from the first value to the last value in the last dense unit. After that, from the remaining dense units, we continue to choose the starting value which is the smallest in the remaining dense units, and expand the values in the current unit and the adjacent dense units, until a non-dense unit is reached. By this way, we can find all the ranges for the numerical attribute values, that is, the numerical attribute values can be discretized. For example, the attribute Age can be discretized as the two ranges [22,24] and [30, 34], and there is only one range [21000, 38000] in the attribute Salary.

After finding all the ranges for a numerical attribute, each attribute value range can be regarded as an item. If the support count of a range is less than the minimum support threshold, then this range can be eliminated. Otherwise, this range is a frequent item. For the above example, suppose the minimum support threshold is 40%. The frequent item for the numerical attributes are (Age,22-24), (Age,30-34) and (Salary,21000-38000). The numerical attribute values also can be encoded according to the range where the attribute values located.

After encoding all the attribute values for the categorical attributes and numerical attributes, our algorithm scans the transaction database TD once to find all the frequent items. For the multidimensional database in Table 5, the corresponding transaction database is shown in Table 6. Suppose the minimum support is set to be 30%., that is, the minimum support count is 30%×8=2.4. Therefore, the frequent items and their associated support counts in Table 6 are (1,1):3, (1,2):5, (2,1):3, (2,2):3, (3,1):4, (3,2):4, (4,1):5, (4,2):3 and (5,1):6.

Table 5. Another multidimensional database

ID	Field 1	Field 2	Field 3	Field 4	Field 5
1	1	1	1	1	1
2	2	2	1	2	1
3	2	2	2	2	1
4	1	1	2	1	1
5	2	3	2	1	2
6	1	1	1	1	1
7	2	3	2	1	2
8	2	2	1	2	1

Table 6. A transaction database transformed from Table 5

T_ID	Itemset
1	(1,1)(2,1)(3,1)(4,1)(5,1)
2	(1,2)(2,2)(3,1)(4,2)(5,1)
3	(1,2)(2,2)(3,2)(4,2)(5, 1)
4	(1,1)(2,1)(3,2)(4,1)(5,1)
5	(1,2)(2,3)(3,2)(4,1)(5,2)
6	(1,1)(2,1)(3,1)(4,1)(5,1)
7	(1,2)(2,3)(3,2)(4,1)(5,2)
8	(1,2)(2,2)(3,1)(4,2)(5,1)

Table 7. The conditional database for frequent item (1,2)

T_ID	Itemset
1	(5,1)(4,1)(3,1)(2,1)
2	(5,1)(3,1)(2,2)(4,2)
3	(5,1)(3,2)(2,2)(4,2)
5	(4,1)(3,2)
7	(4,1)(3,2)
8	(5,1) (3,1)(2,2)(4,2)

For each record in the transaction database TD, our algorithm orders the items in the record according to the supports of the frequent items, that is, order the items in an descending order, and eliminate non-frequent items. For the above example, the order of the frequent items is (5,1), (1,2), (4,1), (3,1), (3,2), (1,1), (2,1), (2,2) and (4,2). For each frequent item, the *conditional database* with this item as prefix can be generated by removing the item from each record containing the item in the transaction database. For the above example, the conditional database for the frequent item (1,2) is shown in Table 7. After generating the conditional database for a length k (k ≥ 1) frequent itemset X, our algorithm continues to recursively find the frequent items from the conditional database, which combine with X to form the length k+1 frequent itemsets.

For example, the frequent items in the conditional database for the frequent 1-itemset (1,2) are (5,1), (4,1), (3,1), (3,2), (2,2) and (4,2). Therefore, the frequent 2-itemsets containing (1,2) are (1,2)(5,1), (1,2)(4,1), (1,2)(3,1), (1,2)(3,2), (1,2)(2,2) and (1,2)(4,2). Our algorithm continues to generate the conditional databases for these frequent 2-itemsets from the conditional database for (1,2), and generate all the frequent 3-itemsets containing (1,2) from these conditional databases. For example, the frequent items in the conditional database for itemset (1,2)(5,1) are (3,1), (2,2) and (4,2), which the frequent 3-itemsets (1,2)(5,1)(3,1), (1,2)(5,1)(2,2) and (1,2)(5,1)(4,2) can be generated.

5 Conclusions

This paper proposes an efficient approach to discover multidimensional frequent patterns in a relational database and automatically discretize numerical-type attributes without experts. Our approach does not need to generate a lot of candidates and scan the original database many times. For a specific target attribute, we can aim at the target attribute to generate conditional database and find all the multidimensional frequent patterns about the target attribute. Many previous approaches for mining multidimensional frequent patterns always need background knowledge to discretize the numerical-type attributes. Our approach uses clustering algorithm to discretize numerical-type attributes without experts. By using the multidimensional frequent patterns, we can promote or recommend the products to the right customers.

References

1. Agrawal, R., et al.: Fast Algorithm for Mining Association Rules. In: Proceedings of International Conference on Very Large Data Bases, pp. 487–499 (1994)
2. Chiang, J., Wu, C.C.: Mining multi-dimensional association rules in multiple database segmentation. In: Proceedings of International Conference on Information Management (2005)
3. El-Hajj, M., Zaiane, O.R.: Non Recursive Generation of Frequent K-itemsets from Frequent Pattern Tree Representation. In: Kambayashi, Y., Mohania, M., Wöß, W. (eds.) DaWaK 2003. LNCS, vol. 2737, pp. 371–380. Springer, Heidelberg (2003)
4. Frawley, W., Piatetsky-Shapiro, G., Matheus, C.: Knowledge Discovery in Databases: An Overview. AI Magazine, 213–228 (1992)
5. Han, J., Pei, J., Yin, Y., Mao, R.: Mining Frequent Patterns without Candidate Generation: A Frequent- Pattern Tree Approach. Data Mining and Knowledge Discovery 8(1), 53–87 (2004)
6. Han, J., Lakshmanan, L.V.S., Ng, R.T.: Constraint-Based, Multidimensional Data Mining. IEEE Computer 32(8), 46–50 (1999)
7. Lee, Y.S., Yen, S.J., Lin, S.S., Liu, Y.C.: Integrating Multidimensional Association Rule Mining into Classification. In: Proceedings of International Conference on Informatics, Cybernetics, and Systems, pp. 831–836 (2003)
8. Park, J.S., Chen, M.S., Yu, P.S.: An Effective Hash-Based Algorithm for Mining Association Rules. Proceedings of ACM SIGMOD 24(2), 175–186 (1995)

9. Tasi, S.M., Chen, C.-M.: Mining interesting association rules from customer databases and transaction databases. Information Systems 29(8), 685–696 (2004)
10. Xu, W., Wang, R.: A Novel Algorithm of Mining Multidimensional Association Rules. In: Proceedings of International Conference on Intelligent Computing, pp. 771–777 (2006)
11. Yen, S.J., Chen, A.L.P.: An Efficient Approach to Discovering Knowledge from Large Databases. In: Proceedings of the International Conference on Parallel and Distributed Information Systems, pp. 8–18 (1996)
12. Yen, S.J., Wang, C.K., Ouyang, L.Y.: A Search Space Algorithm for Mining Frequent Patterns. Journal of Information Science and Engineering (JISE): Special Issue on Technologies and Applications of Artificial Intelligence 28(1), 177–191 (2012)

A Hybrid Cloud for Effective Retrieval from Public Cloud Services

Yi-Hsing Chang[1] and Jheng-Yu Chen[2]

[1] Dept. of Information Management, Southern Taiwan University of Science and Technology,
Tainan County, Taiwan
yhchang@mail.stust.edu.tw
[2] Dept. of Information Management, Southern Taiwan University of Science and Technology,
Tainan County, Taiwan
ma090110@stust.edu.tw

Abstract. For a hybrid cloud, efficient identity authentication and services retrieval are the two most important issues. Therefore, this paper proposes an architecture for a hybrid cloud that solves these two problems. The services requiring a large amount of resources are designed to be intelligent agents and embedded into a private cloud for applications supported by the Google application engine (GAE). When the users acquire the services from the hybrid cloud, these intelligent agents efficiently authenticate the access rights and retrieve the required services for users.

Keywords: Cloud computing, Hybrid Cloud, GAE, OAuth, OpenID.

1 Introduction

Cloud computing has become common in recent years. It offers a cost-effective solution to the problem of how to provide services, data storage and computing power to a growing number of Internet users, without investing large amounts of capital on machinery that must be regularly maintained and upgraded on-site by support staff [1]. From the organization's point of view, cloud computing helps reduce the cost of hardware investment and maintenance.

From the users' point of view, the cloud moves computing and data storage from local devices to remote server pools, so the hardware requirements for local devices are not great. Users use a web browser to access cloud services, such as Google Docs, which is provided by a cloud service provider, using a variety of network devices, such as PC's, notebooks, cell phones, tablets, etc. In the past, Office systems could only be installed on a PC, but now, without the need for software, users can edit or download documents on the cloud, using the browser. In addition, cloud computing also allows interoperability and data transfer between various system platforms, such as Microsoft, Mac, Android, etc.

According to the National Institute of Standards and Technology [2], cloud computing has three service models: Infrastructure as a service (IaaS), Platform of a service

A. Selamat et al. (Eds.): ACIIDS 2013, Part I, LNAI 7802, pp. 61–69, 2013.

(PaaS) and Software as a service (SaaS). It also has the following four deployment models: a public cloud, a private cloud, a hybrid cloud and a community cloud. The applications for the two layers are shown in Table 1.

Table 1. Cloud Applications (self-organized)

	Public cloud	**Private cloud**	**Hybrid cloud**
IaaS	EC2	OpenNebula	EC2 + OpenNebula
PaaS	GAE	Apache	GAE + Apache
SaaS	Google Search	Nutch	Google Search + Nutch

Xu. [3] defined a hybrid cloud as consisting of multiple internal (private) or external (public) clouds. The added complexity of determining how to distribute applications across both private and public clouds is a challenge. Enterprises must strategically leverage deployment models, when a hybrid cloud is established.

This paper examines the use of GAE and existing websites (private cloud) to construct a hybrid cloud, in order to reduce the need for local computing resources, to increase existing website functions and to allow secure access for services between a private cloud and a public cloud.

2 Related Work

2.1 Cloud Computing

The term, cloud computing, was first proposed by Google in 2006. Vouk [4] noted that cloud computing - a relatively recent term, builds on decades of research in virtualization, distributed computing, utility computing and, more recently, networking, web and software services. Therefore, cloud computing is a new concept, rather than a new technology.

Cloud computing has been widely used in many fields in recent years, such as education, manufacturing and e-commerce. In the field of academic research, Doelitzscher, et al. [5] demonstrated how cloud computing allowed the Hochschule Furtwangen University's (HFU) IT department to improve its handling of periodical IT tasks, such as the management of project PCs or lectures, using a standard development environment for programming exercises. Because of its flexibility it is now possible to support a collaboration platform with an on-demand solution with which HFU affiliates are already familiar.

In the field of manufacturing research, Xu[3] demonstrated how cloud computing can transform the traditional manufacturing business model, to align product innovation with business strategy and to create intelligent factory networks that encourage effective collaboration. Manufacturing units are starting to take advantage of cloud computing, because it simply makes good economic sense.

In the field of e-commerce research, Buyya, et al. [6] proposed architecture for the market-oriented allocation of resources within clouds. They also presented a vision for the creation of a global cloud exchange for trading services. In particular, they presented various practical cloud solutions from the market perspective, to demonstrate the potential for the creation of third-party services to enable the successful adoption of cloud computing, such as a meta-negotiation infrastructure for global cloud exchanges that allows high performance delivery of content using storage clouds.

2.2 Google App Engine

The Google App Engine (GAE), a new generation of cloud computing-based web application development platform, enables its users to develop and operate web applications within the Google infrastructure [7]. The GAE was first released as a preview version in April 2008. It is a platform as a service (PaaS) cloud-computing platform for developing and hosting web applications. It supports the Python, Java and Go programming languages and contains the functions, Bitable and Dashboard.

Bitable is a compressed, high performance and proprietary data storage system that uses the Google file system, chubby lock service, SSTable and a few other Google technologies [8]. Bitable is not a traditional relational database, in that it does not support the SQL JOIN syntax. Bitable has the advantage of scalability and it offers good performance [9]. Google Dashboard provides a variety of management and monitoring tools, such as logging, traffic statistics and database management. It allows developers to manage and monitor their applications. The service also allows users to tap into Google's authentication services to identify and authorize users, which eliminates some of the complication in developing a Web application [10].

The GAE allows developers to minimize the work of system implementation and maintenance. Its pricing method is pay-as-you-go. For computing resources, such as storage capacity, bandwidth and CPU, within a certain range, users do not need to pay. If resources must be expanded, users pay a small fee for the use of more computing resources. This pay-as-you-go scenario will revolutionize manufacturing in the same way that the Internet has already revolutionized everyday and business lives [3].

2.3 Authentication

Information security has always been an important issue, especially in the cloud computing age, because all services are accessed via the Internet and mutual authentication is required between services.

Gonza´lez, et al.[11] proposed a solution which they called reverse OAuth, which addresses an identified need for authentication. Reverse OAuth is a useful model for a single sign-on in e-learning systems, given the widespread use of learning management systems.

Hwang & Li [12] suggested the use of a trust-overlay network over multiple data centers to implement a reputation system that establishes trust between service providers and data owners. Data coloring and software watermarking techniques protect shared data objects and widely distributed software modules. These techniques safeguard multi-directional authentication, enable a single sign-on for the cloud and tighten access control for sensitive data, in both public and private clouds.

Shehab & Marouf[13] proposed an extension to the OAuth 2.0 authorization that enables the provision of fine-grained authorization recommendations to users, when granting permission to third party applications. Their experiments using the collected data indicate that the proposed framework enhances user awareness and the privacy related to third-party application authorizations.

3 Problem Description

This paper examines two problems when constructing a hybrid cloud. The first is authentication within the cloud, and the second is the mutual authentication of web services between a private cloud and a pubic cloud. They are described as follows.

3.1 Authentication within the Cloud

Authentication is the first hurdle for users in operating a website function. In most websites, such as Gmail, users can use website functions only after authentication. The operation process is as follows: registration on the website; login to the website using a registered account, password and the operation of website functions by website members. However, a user may have many website passwords. After a period of time, the user may forget the ID and passwords for some websites. However, a user who uses the same ID and password for all websites is at risk. If the password is stolen by a malicious websites, personal privacy is compromised. The probability of these problems occurring is higher in a cloud environment, because users only need to operate software on a PC, without the need to perform authentication on the Internet, as in the past.

This paper, constructs a hybrid cloud based on Moodle and the Google web services. Moodle has its own user database and Google also has its own user management mechanism. Since the two websites' verification use different methods, the user must login to the websites separately. In order to address this problem, an intelligent agent solves the problem of user authentication, using GAE APIs.

3.2 Mutual Authentication between Private Cloud and Public Cloud Web Services

The proposed hybrid cloud architecture embeds web services developed on GAE in the Moodle. The problem is the safe access of services between the two websites. For example, if one website provides a text edition service and the other website provides a data storage service, then if the user wants to directly edit a file stored in the in the text edition service website of the other website, an ID and password for the data storage website must be provided. However, this is not safe, because the ID and password could be stolen.

This paper uses the OAuth API provided by the GAE as an authentication channel between the two websites. The user can allow a third party to access network application programs on their behalf, without sharing the authentication (ID and password) with the third party [14].

4 Proposed Architecture

In order to conserve the computing resources of the local server and to allow retrieval from public cloud services, a hybrid cloud is proposed, as shown in Fig. 1. This contains a private cloud, a public cloud and a database. The design integrates the private cloud and public cloud services into a hybrid, using intelligent agents. The description is as follows:

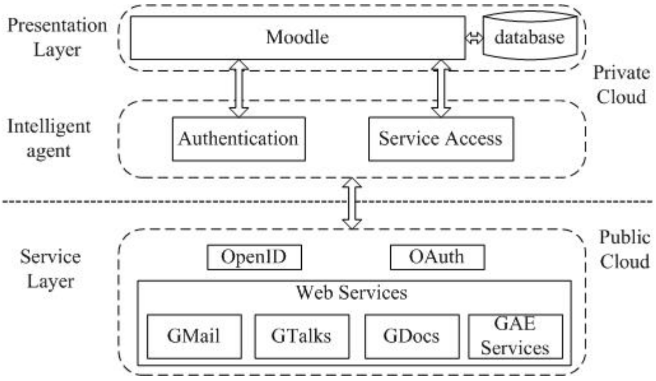


Fig. 1. Hybrid cloud framework

4.1 Private Cloud

This consists of the Moodle, a database and intelligent agents.

Moodle. Moodle is a Course Management System (CMS), also known as a Learning Management System (LMS). It is a free web application that educators can use to create effective online learning sites. The Moodle is the e-learning portal and users can use the functions of Moodle, Google and the web services developed by the GAE after they login to the system once.

Database. This contains the Moodle database and a web service database.

- Moodle database: This database is automatically created when Moodle is installed and stores the required data, such as user information, course information and learning portfolio.
- Web service database: The data generated and collected by web services, such as usage time, the frequency of use and learning effectiveness, are stored in this database.

Intelligent Agent. There are two intelligent agents, one for authentication and one for service access.

- Authentication intelligent agent: This agent addresses the problem posed by the use of different methods of user authentication for Moodle and Google. The GAE OpenID API is used to design an agent that allows users to sign in to two different websites at the same time using only a Google account. They do not need to log in to both, separately.
- Service access intelligent agent: The GAE OAuth API is used to allow this agent to overcome the problem of service retrieval between two websites. For example, if a user wants to use GDocs to edit a file stored in the Moodle database, the file must be downloaded and then they must login to GDocs and upload the file. Using the service access intelligent agent, users simply use the GDocs supported in Moodle in one step, to complete this action.

4.2 Public Cloud

This contains the following three elements: OpenID, OAuth and Web services.

OpenID. App Engine applications authenticate users using any one of three methods: Google accounts, accounts on individual Google Apps domains, or OpenID identifiers [15]. Fig. 2 shows the process flow for authentication.

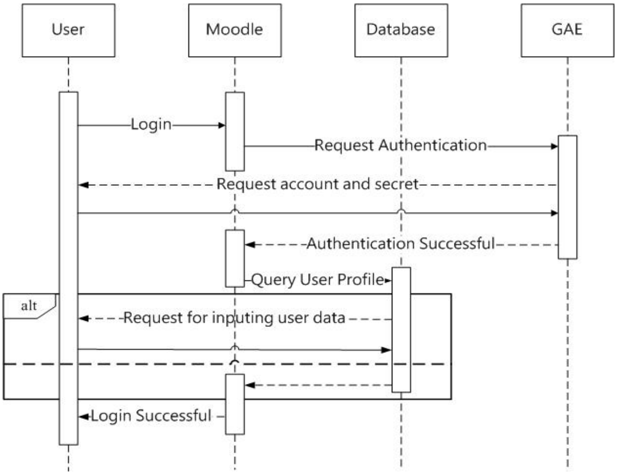


Fig. 2. The process flow for Authentication

When a user connects to Moodle and requests login, Moodle calls the authentication agent to send the request to Google. The GAE determines whether the user has logged in. If not, it returns the webpage to the user, asking for a password. When the authentication by GAE is successful, the GAE returns the user's ID to the intelligent agent. The authentication agent then informs the Moodle of a successful authentication. If it finds that the user has not been logged in before, Moodle returns the webpage, asking for the user's personal information. Once this personal information has been input, Moodle returns a successful login message to the user.

OAuth. The OAuth framework provides a mechanism by which third party service providers can access end-user resources, without releasing the user's access credentials to the service provider [13]. This architecture allows mutual authentication between the Moodle and the GAE web services. For example, the OAuth framework might allow the sharing of photographs from a primary web-based photo sharing website, so that a third-party photo printing service can access the authorized photographs [16].

Web services. This paper uses the Moodle as a portal website and users can directly acquire the web services in Google from Moodle. Fig. 3 shows the process flow for web service retrieval. Firstly, users request the web services provided by Moodle. The service access intelligent agent requests Google OAuth to verify whether Moodle has the authority to access the web services. The identification of users is processed using Google OpenID. If a user does not login, the system returns the login prompt page to the user. When the user logs in successfully, the Moodle obtains service authorization. The Moodle then determines whether the user has registered, by searching the database. If the user has registered previously, the user can use the requested web services.

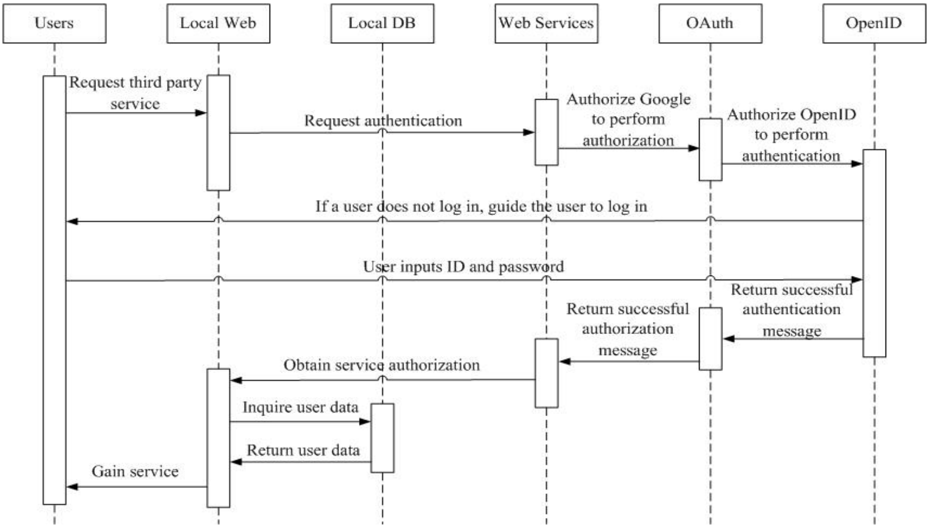

Users
Local Web
Local DB
Web Services
OAuth
OpenID
Request third party service
Request authentication
Authorize Google to perform authorization
Authorize OpenID to perform authentication
If a user does not log in, guide the user to log in
User inputs ID and password
Return successful authentication message
Return successful authorization message
Obtain service authorization
Inquire user data
Return user data
Gain service

Fig. 3. The process flow for web services retrieval

4.3 Features of the Proposed Hybrid Cloud

Combining Moodle, cloud OpenID and OAuth has the following 4 advantages:

- It reduces the consumption of local computing resources: Moodle has its own authentication mechanism, but CPU resources are consumed because it must query a back-end database for each user login. If many users login simultaneously, the CPU will be too busy and a computer crash occurs. The proposed hybrid cloud can place these resource-consuming programs in another cloud, to reduce the local burden.

- Provision of a single sign-on: Users use a Google ID to login to Moodle and are not required to remember many passwords. If a user uses the same ID and password for all websites, the ID and password could be stolen by malicious websites.
- Prevention of a brute force attack: The main problem for websites is security and the most common mode of attack is a password breaker. In this paper, the Google authentication mechanism is used and this provides good security.
- The allowance of mutual authentication between the Moodle and GAE web services: The problem of mutual authentication between Moodle and GAE web services is addressed using the OAuth framework, so users can access the web services provided in the hybrid cloud.

5 Conclusions

As the concept of cloud computing has emerged, website service providers, as well as governments, enterprises and schools, have invested in their own cloud services. Cloud computing has become mainstream. However, cloud computing requires large amounts of software and hardware and has high maintenance costs. Small organizations may not be able to bear these costs, so a framework is proposed for efficient identity authentication and services retrieval, using a hybrid cloud. The Moodle and Google web services are integrated together by the intelligent agents within the hybrid cloud.

The proposed hybrid cloud has the following 4 advantages: a reduction in the consumption of local computing resources, the provision of a single sign-on, the prevention of a brute force attack, the allowance of mutual authentication between the Moodle and GAE web services. Future studies will analyze the security of this architecture, determine how to synchronize data in the local database and the GAE Bigtable and improve database access functions.

Acknowledgment. This study is supported in part by the National Science Council of Republic of China under the contract number NSC 101-2511-S-218 -008 -.

References

1. Eason, M., El-Seoud, S.A., Wyne, M.F.: Cloud Computing Based E-Learning System. International Journal of Distance Education Technologies 8(2), 58–71 (2010)
2. Mell, P., Grance, T.: The NIST Definition of Cloud Computing Ver. 15, 10-7-09. NIST, Information Technology Lab (2009)
3. Xu, X.: From cloud computing to cloud manufacturing. Robotics and Computer-Integrated Manufacturing 28(1), 75–86 (2012)

4. Vouk, M.A.: Cloud Computing Issues Research and Implementations. Journal of Computing and Information Technology 16, 235–246 (2008)
5. Doelitzscher, F., Sulistio, A., Reich, C., Kuijs, H., Wolf, D.: Private Cloud for Collaboration and e-Learning Services:from IaaS to SaaS. Computing 91, 23–42 (2011)
6. Buyya, R., Yeo, C.S., Venugopal, S., Broberg, J., Brandic, I.: Cloud computing and emerging IT platforms: Vision, hype, and reality for delivering computing as the 5th utility. Future Generation Computer Systems 25(6), 599–616 (2009)
7. Xu, H.X.: Preliminary Study on the Influence of Cloud Computing on Education. Computer Knowledge and Technology 5, 2690–2692 (2009)
8. BigTable, http://en.wikipedia.org/wiki/BigTable
9. BigTable, http://zh.wikipedia.org/wiki/BigTable
10. Bedra, A.: Getting Started with Google App Engine and Clojure. IEEE Internet Computing 14(4), 85–88 (2010)
11. González, J.F., Rodríguez, M.C., Nistal, M.L., Rifón, L.A.: Reverse OAuth: A solution to achieve delegated authorizations in single sign-on e-learning systems. Computers & Security 28(8), 843–856 (2009)
12. Hwang, K., Li, D.: Trusted Cloud Computing with Secure Resources and Data Coloring. Internet Computing 14(5), 14–22 (2010)
13. Shehab, M., Marouf, S.: Recommendation Models for Open Authorization. IEEE Transactions On Dependable And Secure Computing 9(4), 583–596 (2012)
14. OAuth for java,
https://developers.google.com/appengine/docs/java/oauth/
15. Users Java API Overviews,
https://developers.google.com/appengine/docs/java/users/overview
16. Bin, W., Yuan, H.H., Xi, L.X., Min, X.J.: Open Identity Management Framework for Saas Ecosystem. In: Proc. IEEE Int'l Conf. e-Business Eng (ICEBE 2009), pp. 512–517. IEEE Computer Society, Washington (2009)

A New Method for Generating the Chinese News Summary Based on Fuzzy Reasoning and Domain Ontology

Shyi-Ming Chen and Ming-Hung Huang

Department of Computer Science and Information Engineering,
National Taiwan University of Science and Technology,
Taipei, Taiwan

Abstract. This paper presents a new method for automatically generating the Chinese weather news summary based on fuzzy reasoning and the domain ontology, where the weather ontology, the time ontology and the geography ontology are predefined by domain experts. The summary is composed of candidate sentences which have higher scores, where the experimental data are adopted from the Chinese weather news website of Taiwan. The experimental results show that the proposed method outperforms the methods presented in [9] and [10] for automatically generating the Chinese news summary.

Keywords: Chinese news summary, domain ontology, fuzzy reasoning, fuzzy sets.

1 Introduction

Text summarization [2], [9], [10], [12], [13] is a task of creating a document from one or more textual sources that are smaller in size, but hold some or most of the information contained in the original sources [7].

Many ontological applications have been presented in various domains [1], [3], [4], [6], [11], [14]. The concept of ontology comes from philosophy, and refers to the subject of existence. It can describe the existence of instances or things in the real-world. We can treat ontology as a formal explicit description of concepts in a domain, where properties of each concept describing various features and attributes of concepts and various restrictions on slots.

In this paper, we present a new method for automatically generating the Chinese weather news summary based on fuzzy reasoning and the domain ontology, where the weather ontology, the time ontology and the geography ontology are predefined by domain experts. The summary is composed of candidate sentences which have higher scores, where the experimental data are adopted from the Chinese weather news website of Taiwan. The experimental results show that the proposed method outperforms the methods presented in [9] and [10] for automatically generating the Chinese news summary.

A. Selamat et al. (Eds.): ACIIDS 2013, Part I, LNAI 7802, pp. 70–78, 2013.

2 Fuzzy Rules and Fuzzy Reasoning Techniques

A fuzzy set A in the universe of discourse U, where $U = \{x_1, x_2, \ldots, x_n\}$, can be represented as follows [16]:

$$A = \mu_A(x_1)/x_1 + \mu_A(x_2)/x_2 + \cdots + \mu_A(x_n)/x_n, \tag{1}$$

where μ_A is the membership function of the fuzzy set A, $\mu_A(x_i)$ denotes the degree of membership of element x_i in the fuzzy set A, $\mu_A(x_i) \in [0, 1]$ and $1 \le i \le n$. Fig. 1 shows a triangular fuzzy set A with the membership function μ_A parameterized by a triplet (a, b, c), where $A = (a, b, c)$ and a, b and c are called the left vertex, the center vertex and the right vertex of the fuzzy set A, respectively.

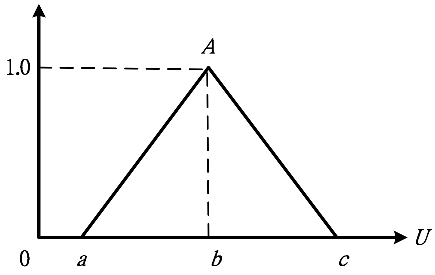


Fig. 1. A triangular fuzzy set

Let A and B be two fuzzy sets in the universe of the discourse U and let the membership function of the fuzzy sets A and B be μ_A and μ_B, respectively. The intersection between the fuzzy sets A and B is defined as follows [16]:

$$\mu_{A \cap B}(x) = \min\{\mu_A(x), \mu_B(x)\}, \forall x \in U. \tag{2}$$

The union between the fuzzy sets A and B is defined as follows [16]:

$$\mu_{A \cup B}(x) = \max\{\mu_A(x), \mu_B(x)\}, \forall x \in U. \tag{3}$$

Let us consider the following fuzzy rules:

Rule 1: IF X is A_1 and Y is B_1 THEN Z is C_1,

Rule 2: IF X is A_2 and Y is B_2 THEN Z is C_2,

where X, Y and Z are linguistic variables, A_1, A_2, B_1, B_2, C_1 and C_2 are linguistic values represented by fuzzy sets. According to [8], Mamdani's Max-Min operations for fuzzy reasoning are shown in Fig. 2, where the system performs the defuzzification operation to get a crisp value z of the fuzzy reasoning result based on the center of gravity (COG) defuzzification method, shown as follows:

$$z = \frac{\sum_{i=1}^{k} \mu_C(z_i) z_i}{\sum_{i=1}^{k} \mu_C(z_i)}, \tag{4}$$

where μ_C is the membership function of the fuzzy set C, $\mu_C(z_i)$ denotes the degree of membership of element z_i belonging to the fuzzy set C and $1 \leq i \leq k$.

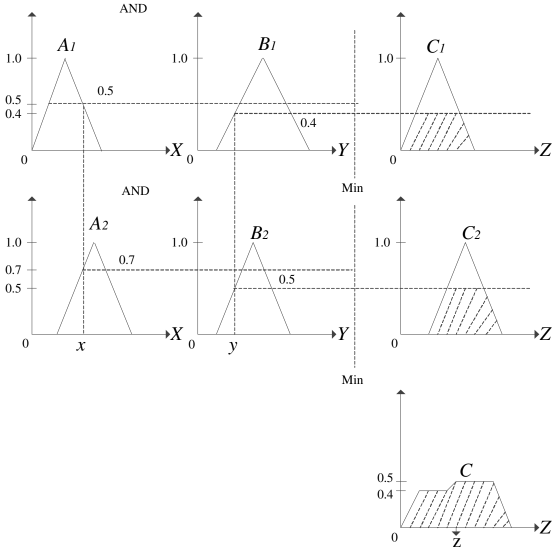


Fig. 2. Mamdani's Max-Min operations for fuzzy reasoning [8]

3 A New Method for News Summarization Generation Based on Fuzzy Reasoning and the Domain Ontology

In this section, we present a new method for news summarization generation based on fuzzy reasoning and the domain ontology, as shown in Fig. 3,

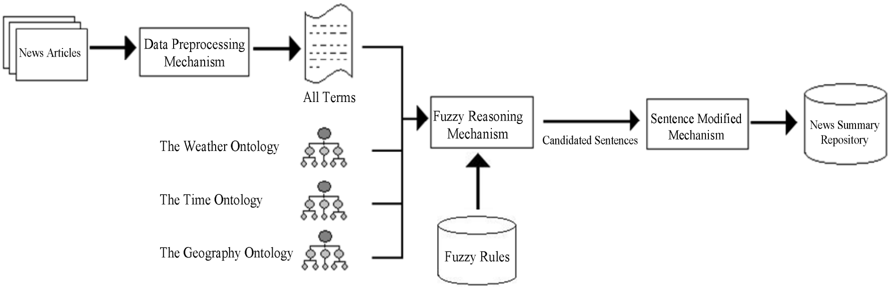


Fig. 3. The proposed news summarization generation method

In the following, we describe the data preprocessing mechanism shown in Fig. 3, where “terms” are minimum units in an article. A number of terms constitute a sentence. A sentence or a few sentences can express a complete meaning of an article. In the proposed method, all Chinese weather news articles are retrieved from the Internet and are stored into a text corpus. In order to further parsing and analyzing the meaning of a sentence, we slice each sentence into a set of terms. There are two steps in the data preprocessing mechanism. First, we divide an article into a set of sentences by detecting the punctuation (i.e. the symbols “ ， ” , ” 。 ” and ”?”). Then, we slice all the sentences into a set of terms. In recent years, some researchers have focused on the research topic of how to divide sentences into terms correctly and detect unknown terms. In this paper, we apply the program provided by the Chinese Knowledge Information Processing (CKIP) Group [15] to divide a sentence into a set of terms, where each term is associated with a part-of-speech (POS) tag. After this process, we can get a set of terms with POS tags. In the data preprocessing algorithm, the total number of news is m, N_i denotes the *ith* news in the text corpus, the news N_i is sliced into a set S_i of sentences, where S_{ij} denotes the jth sentence in the ith news N_i, and u_i denotes the number of sentences in the ith news N_i, where $1 \le i \le m$. Then, each sentence S_{ij} is segmented into a set T_{ij} of terms, where T_{ijk} denotes the kth term in the jth sentence of the ith news N_i, where $1 \le i \le m$. Let v_{ij} denote the number of terms in the jth sentence of the ith news N_i, where $1 \le i \le m$. In order to evaluate the performance of the proposed method, three experts predefined the gold summary GS_N_i, for each news N_i and stored the gold summary GS_N_i of the ith news to the gold summary text corpus, where $1 \le i \le m$.

The news summarization system shown in Fig. 3 will assign a score to each sentence. A higher score assigned to a sentence implies a higher possibility for the sentence to be selected into the news summary. In the following, we apply the fuzzy reasoning techniques to compute the score for each sentence. There are three input linguistic variables of the fuzzy reasoning mechanism shown in Fig. 3, i.e., “the Depth”, “the Width” and “the Frequency”. The input linguistic variable “the Depth” denotes the average degree of the depth of terms in a sentence, where the degree of depth of a term is calculated by the length of the path from the root node to the concept node. There are three linguistic terms “Low”, “Medium” and “High” for the linguistic variable “the Depth”. A smaller length of path implies a higher chance for a sentence to be selected into the news summary. Because the news summarization system tends to extract the abstract information and ignore the details, if a concept has a smaller distance from the root to the concept, then it will have a higher score. The linguistic variable “the Width” denotes how many domain ontologies are covered by a concept set in a sentence, where the sets of concepts are retrieved from the domain ontologies. There are three linguistic terms “Low”, “Medium” and “High” for the linguistic variable “the Width”. If a sentence is covered by more ontologies, then it means that there is more relevant and crucial information in this sentence.

We also consider that each term contributes to the news summary, where the terms are retrieved from the original news. The linguistic variable “the Frequency” denotes the significance of a term toward the news summary. There are three linguistic terms “Low”, “Medium” and “High” for the linguistic variable “the Frequency”. The frequency “*frequency*(t)” of a term t is defined as follows:

$$Frequency(t) = \frac{I(t)}{I(t)'}, \tag{5}$$

where *I*(*t*) and *I*(*t*)' are defined as follows [12]:

$$I(t) = -\frac{\log(P(t))}{\log(f+1)}, \tag{6}$$

$$I(t)' = -\frac{\log(P(t)')}{\log(f'+1)}, \tag{7}$$

P(*t*) represents the probability of term *t* in the text corpus and *P*(*t*)' denotes the probability of term t in the gold summary, where

$$P(t) = \frac{e+1}{f+1}, \tag{8}$$

$$P(t)' = \frac{e'+1}{f'+1}, \tag{9}$$

f denotes the total number of terms in the text corpus and e denotes the frequency of term *t* in the text corpus. On the other hand, *e*' denotes the total number of terms in the gold summary corpus and *e* denotes the frequency of term *t* in the gold summary corpus. If *I*(*t*) is larger than *I*(*t*)', then it means that term *t* has less information. On the other hand, if *I*(*t*)' is larger than *I*(*t*) then it means that term *t* has more information toward the gold summary.

Table 1 shows the corresponding triangular fuzzy sets for the linguistic terms of the linguistic variables "the Depth", "the Width", "the Frequency" and "the Score" predefined by domain experts, respectively.

Table 1. Linguistic terms of the linguistic variables and their corresponding triangular fuzzy sets

Linguistic Variables	Linguistic Terms	Triangular Fuzzy Sets (*a*, *b*, *c*)
Depth	Low	(0, 0, 0.5)
	Medium	(0, 0.5, 1)
	High	(0.5, 1, 1)
Width	Low	(0, 0, 1.5)
	Medium	(0, 1.5, 3)
	High	(1.5, 3, 3)
Frequency	Low	(0, 0, 0.5)
	Medium	(0, 0.5, 1)
	High	(0.5, 1, 1)
Score	Very low	(0, 0, 2)
	Low	(0, 2, 4)
	Medium	(2, 4, 6)
	High	(4, 6, 8)
	Very high	(6, 8, 8)

In the proposed method, we use the 27 fuzzy rules shown in Table 2 to infer the degree of membership of the output linguistic variable "the Score".

Table 2. Fuzzy rule matrix to infer the score

Depth Width Frequency	Low			Medium			High		
	Low	Medium	High	Low	Medium	High	Low	Medium	High
Low	VL	VL	VL	L	L	M	L	M	M
Medium	VL	L	M	L	M	H	M	H	H
High	L	L	M	M	H	H	M	H	VH

In the fuzzy reasoning mechanism shown in Fig. 3, the score $Score_{ij}$ of the jth sentence in the ith news N_i is represented by a triplet ($Depth_{ij}$, $Width_{ij}$, $Frequency_{ij}$), where $Score_{ij}$ denotes the score of the jth sentence of the ith news N_i, $Depth_{ij}$ denotes the degree of the depth of the jth sentence in the ith news N_i, $Width_{ij}$ denotes the degree of width of the jth sentence in the ith news N_i, and $1 \leq i \leq m$. The definition of $Width_{ij}$ is shown as follows:

$$Width_{ij} = O_{1ij} + O_{2ij} + \cdots + O_{nij}, \tag{10}$$

where O_{pij} is a two dimensional boolean array to record the pth ontology used in the jth sentence of the ith news N_i, p denotes the pth ontology, $1 \leq i \leq m$, and $1 \leq p \leq$ n. The initial value of $O_{pij} = 0$ means that the pth ontology is not used in the jth sentence of the ith news N_i, where $1 \leq i \leq m$. On the other hand, $O_{pij} = 1$ means that the pth ontology is used in the jth sentence of the ith news N_i, where $1 \leq i \leq m$. After performing the fuzzy reasoning algorithm, we can obtain the score $Score_{ij}$. of the jth sentence S_{ij} in the ith news N_i, where $1 \leq i \leq m$. A threshold value ε is used to identify the significance of each sentence, where the threshold value ε is given by the expert and $4 \leq \varepsilon \leq 8$. If the value of $Score_{ij}$ is larger than or equal to ε, then the sentence S_{ij} is a candidate sentence. After this process, we get a set of candidate sentences which have higher scores as a news summary. But sometime the semantic of the sentences in the news summary may be confusing. In this situation, we modify the sentence by removing the relevant conjunction term "but" according to the category "*C*" of the POS tags provided by the CKIP group [15]. Therefore, the sentence modifying process detects the relevant conjunction term "but" and removes it from the sentence.

4 Experimental Results

In this paper, the experimental data is adopted from the Chinese weather news website (http://www.chinatimes.com) in Taiwan. The precision and the recall are defined as follows [12]:

$$Precision = \frac{n}{m}, \tag{11}$$

$$Recall = \frac{n}{o}, \tag{12}$$

where *m* denotes the number of sentences selected by the domain experts, *n* denotes the number of sentences selected by our summarization system and selected by the domain experts, and *o* denotes the total number of sentences selected by the summarization system. In this paper, we choose 180 Chinese weather news from the Chinese weather news website (http://www.chinatimes.com) and three domain experts made this experiment.

In the following, we compare the performance of the proposed method with the ones of Lee, Chen and Jain's method [9] and Lee, Jian and Huang's method [10], as shown from Table 3 to Table 8.

Table 3. A comparison of the average precisions for different methods for the typhoon events of the 20 news for different years

Methods Average Precision Years	Lee, Chen and Jain's Method [9]	Lee, Jain and Huang's Method [10]	The Proposed Method
2002	0.66	0.43	0.79
2003	0.55	0.35	0.78
2004	0.61	0.41	0.75

Table 4. A comparison of the average precisions for different methods for the cold-current events of the 20 news for different years.

Methods Average Precision Years	Lee, Chen and Jain's Method [9]	Lee, Jain and Huang's Method [10]	The Proposed Method
2002	0.50	0.44	0.78
2003	0.51	0.44	0.77
2004	0.39	0.22	0.80

Table 5. A comparison of the average precisions for different methods for the rain events of the 20 news for different years

Methods Average Precision Years	Lee, Chen and Jain's Method [9]	Lee, Jain and Huang's Method [10]	The Proposed Method
2002	0.59	0.59	0.79
2003	0.33	0.32	0.81
2004	0.70	0.75	0.76

Table 6. A comparison of the average recalls for different methods for the typhoon events of the 20 news for different years

Methods Average Recall Years	Lee, Chen and Jain's Method [9]	Lee, Jain and Huang's Method [10]	The Proposed Method
2002	0.51	0.46	0.57
2003	0.58	0.38	0.59
2004	0.53	0.35	0.55

Table 7. A comparison of the average precisions for different methods for the cold-current events of the 20 news for different years

Methods Average Recall Years	Lee, Chen and Jain's Method [9]	Lee, Jain and Huang's Method [10]	The Proposed Method
2002	0.53	0.31	0.55
2003	0.75	0.30	0.54
2004	0.63	0.23	0.58

Table 8. A comparison of the average recalls for different methods for the rain events of the 20 news for different years

Methods Average Recall News Number	Lee, Chen and Jain's Method [9]	Lee, Jain and Huang's Method [10]	The Proposed Method
2002	0.87	0.22	0.60
2003	0.76	0.23	0.60
2004	0.75	0.21	0.57

5 Conclusions

We have presented a new method for automatically generating the Chinese weather news summary based on fuzzy reasoning and the domain ontology, where the weather ontology, the time ontology and the geography ontology are predefined by domain experts. The experimental results show that the proposed method outperforms the methods presented in [9] and [10] for automatically generating the Chinese news summary.

Acknowledgements. This work was supported in part by the National Science Council, Republic of China, under Grant NSC 100-2221-E-011-118-MY2.

References

1. Adnan, M., Warren, J., Orr, M.: Ontology Based Semantic Recommendations for Discharge Summary Medication Information for Patients. In: Proceedings of the 2010 IEEE International Symposium on Computer-Based Medical Systems, pp. 456–461 (2010)
2. Chen, P., Verma, R.: A Query-Based Medical Information Summarization System Using Ontology Knowledge. In: Proceedings of the 2006 IEEE Symposium on Computer-Based Medical Systems, pp. 37–42 (2006)
3. Chen, R.C., Huang, Y.H., Bau, C.T., Chen, S.M.: An Recommendation System Based on Domain Ontology and SWRL for Anti-Diabetic Drugs Selection. Expert System with Application 39(4), 3995–4005 (2012)
4. Fensel, D., Harmelen, F.V., Horrocks, I., McGuinness, D.L., Schneider, P.F.P.: OIL: An Ontology Infrastructure for the Semantic Web. IEEE Intelligent Systems 16(2), 38–45 (2001)
5. Gruber, T.R.: A Translation Approach to Portable Ontology Specifications. Knowledge Acquisition 5(2), 199–220 (1993)
6. Guarino, N.: Understanding, Building and Using Ontologies. International Journal of Human-Computer Studies 46(2-3), 293–310 (1997)
7. Hennig, L., Umbrath, W., Wetzker, R.: An Ontology-Based Approach to Text Summarization. In: Proceedings of the 2008 IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology, pp. 291–294 (2008)
8. Lee, C.C.: Fuzzy Logic in Control Systems: Fuzzy Logic Controller-Part II. IEEE Transactions on Systems, Man and Cybernetics 20(2), 419–435 (1990)
9. Lee, C.S., Chen, Y.J., Jain, Z.W.: Ontology-Based Fuzzy Event Extraction Agent for Chinese e-News Summarization. Expert Systems with Applications 25(3), 431–447 (2003)
10. Lee, C.S., Jian, Z.W., Huang, L.K.: A Fuzzy Ontology and Its Application to News Summarization. IEEE Transactions on Systems, Man, and Cybernetics-Part B: Cybernetics 35(5), 859–880 (2005)
11. Lee, C.S., Wang, M.H., Hagras, H.: A Type-2 Fuzzy Ontology and Its Application to Personal Diabetic-Diet Recommendation. IEEE Transactions on Fuzzy Systems 18(2), 374–395 (2010)
12. Liu, X.Y., Zhou, Y.M., Zheng, R.S.: Measuring Semantic Similarity within Sentences. In: Proceedings of the 2008 International Conference on Machine Learning and Cybernetics, pp. 2558–2562 (2008)
13. Mohamed, A.A.: Generating User-Focused, Content-Based Summaries for Multi-Document Using Document Graphs. In: Proceedings of the 2005 IEEE International Symposium on Signal Processing and Information Technology, pp. 675–679 (2005)
14. Noy, N.F., McGuinnes, D.L.: Ontology Development 101: A Guide to Creating Your First Ontology. Stanford Knowledge Systems Laboratory Technical Report KSL-01-05 and Stanford Medical Informatics Technical Report SMI-2001-0880 (2001)
15. Chinese Knowledge Information Processing Group. Academic Sinica, Taipei, Taiwan, R. O. C., `http://godel.iss.sinica.edu.tw/CKIP/`
16. Zadeh, L.A.: Fuzzy Sets. Information and Control 8, 338–353 (1965)

Hybrid PSO and GA for Neural Network Evolutionary in Monthly Rainfall Forecasting

Linli Jiang[1] and Jiansheng Wu[1,2]

[1] Department of Mathematics and Computer, Liuzhou Teacher College,
Liuzhou, 545004, Guangxi, P.R. China
jll200@163.com

[2] School of Information Engineering, Wuhan University of Technology
Wuhan, 430070, Hubei, P.R. China
wjsh2002168@163.com

Abstract. Accurate and timely weather forecasting is a major challenge for the scientific community in hydrological research such as river training works and design of flood warning systems. Neural Network (NN) is a popular regression method in rainfall predictive modeling. This paper investigates the effectiveness of the hybrid Particle Swarm Optimization (PSO) and Genetic Algorithm (GA) evolved neural network for rainfall forecasting and its application to predict the monthly rainfall in a catchment located in a subtropical monsoon climate in Guilin of China. Our methodology adopts a hybrid Particle Swarm Optimization (PSO) and Genetic Algorithm (GA) for the automatic design of NN by evolving to the optimal network configuration(s) within an architecture space, namely PSOGA–NN. The PSO is carried out as a main frame of this hybrid algorithm while GA is used as a local search strategy to help PSO jump out of local optima and avoid sinking into the local optimal solution early. The proposed technique is applied over rainfall forcasting to test its generalization capability as well as to make comparative evaluations with the several competing techniques, such as GA–NN, PSO–NN and NN. The experimental results show that the GAPSO–NN evolves to optimum or near–optimum networks in general and has a superior generalization capability with the lowest prediction error values in rainfall forecasting. Experimental results reveal that the predictions using the GAPSO–NN approach can significantly improve the rainfall forecasting accuracy.

Keywords: Neural Network, Particle Swarm Optimization, Genetic Algorithm, Rainfall Forecasting.

1 Introduction

Accurate and timely rainfall forecasting is one of the most difficult and important processes of the hydrologic cycle due to the complexity of the physical processes involved and the variability of rainfall in space and time [1, 2]. Although a physically–based approach for rainfall forecasting has several advantages, it is

A. Selamat et al. (Eds.): ACIIDS 2013, Part I, LNAI 7802, pp. 79–88, 2013.

not a feasible alternative in most cases because it involves many variables which are interconnected in a very complicated way, and and the volume of rainfall calculation require sophisticated mathematical tool [3, 4]. Recurrently, Neural Networks (NN), which emulate the parallel distributed processing of the human nervous system, have proven to be very successful in forecasting rainfall data.

Three–layer feedforward NN have been commonly used in modeling forecasting modelling, in which the connection weight training is normally completed by a back–propagation (BP) learning algorithm, and it can approximate arbitrary complex function with any precision [5, 6]. Despite its popularity as an optimization tool for neural network training, the BP algorithm also has several drawbacks. For instance, the performance of the network learning is strictly dependent on the shape of the error surface, values of the initial connection weights, network architecture, training data collection, and different types of activation functions, so that the convergence to the global optimum is not guaranteed. These limitations cause the backpropagation neural network (BPNN) to become inconsistent and unpredictable on rainfall forecasting applications [8, 9].

Evolutionary algorithms (EA) and Swarm Intelligence (SI) algorithms have been widely used in the last few years for training and/or automatically designing neural networks. Kiranyaz S., Ince T. and Yildirim A. et al. have propose a novel multi–dimensional Particle Swarm Optimization (MD–PSO) technique for the automatic design of Neural Networks (NN) by evolving to the optimal network configuration(s) within an architecture space. The experimental results show that the MD–PSO evolves to optimum or near–optimum networks in general and has a superior generalization capability [10]. Wu J., Jin L. and Liu M. have presented that network architecture and connection weights of NN are evolved by PSO method. When evolutionary was end, the appropriate network architecture and connection weights were fed into back–propagation (BP) networks and ensemble strategy is carried out by simple averaging for rainfall forecasting [11]. Sedki, A., Ouazar, D. and Mazoudi, E. E. have designed and optimized the neural network by a real coded GA strategy and hybrid with a BP algorithm for rainfall–runoff forecasting. The paper results showed that the GA-based neural network model gives superior predictions. The well–trained neural network can be used as a useful tool for runoff forecasting [7].

In order to overcome the local optimum problem, PSO has been combined with GA to find a better optimal NN for rainfall modelling. Thus, this paper proposes an evolutionary algorithm for design the network architecture and connection weights based on a hybrid of genetic algorithm (GA) and particle swarm optimization algorithm (PSO), namely PSOGA–NN. The evolved NN is used to establish rainfall forecasting model.

The rest of this study is organized as follows. Section 2 describes the proposed PSOGA–NN, ideas and procedures. Next in Section 3, the application the PSOGA–NN for monthly rainfall forecasting at Guilin of Guangxi is presented and compared with BP–NN and pure GA–NN using the same observed data. Discussions are presented in Section 4 and conclusions are drawn in the final Section.

2 The Developed PSOGA–NN Approach

2.1 Genetic Algorithm and Particle Swarm Optimization

Genetic algorithm is an adaptive optimization technique developed by Holland based on natural evolution and survival of the fittest, and works on a population of individuals, which has good global search characteristics and local optimizing algorithm (LOA) has good local search characteristics [12].

PSO is a stochastic, population–based optimization algorithm introduced in 1995 by James Kennedy and Russell C. Eberhart [13]. PSO is initialized with a population of random solutions. Its developmentwas based on observations of the social behavior of animals such as bird flocking, fish schooling, and swarm theory. Each individual in PSO is assigned with a randomized velocity according to its own and its companions' flying experiences, and the individuals, called particles, are then flown through hyperspace.

Hybridization of evolutionary algorithms with local search has been investigated in many studies [14–16]. Such a hybrid is often referred to as a memetic algorithm. In contrast to memetic algorithm, we will combine two global optimization algorithms, i.e., GA and PSO, for design the network architecture and connection weights in this paper.

2.2 Neural Network Rainfall Forecasting Model

NN is widely recognised for their ability to approximate complicated non–linear relationships and to estimate underlying trends, even when substantial noise is present in the data through the application of many non–linear processing units called neurons [5, 17, 18]. In general, the backpropagation method is used for training multilayer perception neural network. During the design and training of an NN, there are a number of different parameters that must be determined ahead, such as, the number of layers, the number of neurons per layer, the number of training iterations, and so on. There is no general and explicit method for choosing these parameters. In the practical application, researchers determine appropriate network architecture and values of different parameters with trial and error due to short of prior knowledge [6, 11, 19]. Hence, proposing an approach is necessary to find the optimum combination of parameters to affect the performance of NN for rainfall forecasting.

2.3 Evolving Neural Network Using Hybrid PSOGA

In this paper, a three layer back propagation network (Fig. 1) is evolved to predict model, where the transfer functions in hidden and output layer are sigmoid and linear, respectively.

Each the output neurons output is calculated using Equation (2)

$$\hat{y}_i = \sum_{i=1}^{m}(sigmoid(\sum_{i=1}^{n} x_i \omega_{ij} + \omega_{j0}))\omega_{jk} + \omega_{k0} \tag{1}$$

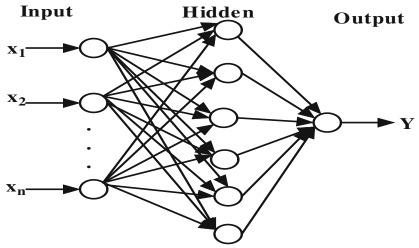


Fig. 1. Architecture of the neural network model used in this paper

where x_i is is the value of the input variable, ω_{ij} and ω_{jk} are connection weights between the input and hidden neuron, and between the hidden neuron and output neuron. ω_{j0} are ω_{k0} the threshold (or bias) for the ith and kth neuron, respectively, and i, j and k are the number of neurons for the layers, respectively.

Suppose we are given training data $(x_i, y_i)_{i=1}^{n}$, where $x_i \in R^n$ is the input vector; y_i is the output value and n is the total number of data dimension. The fitness function is defined as follows:

$$f(\omega) = 1/[1 + \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y_i})^2] \tag{2}$$

The main idea of hybrid PSOGA algorithm is to integrate the GA operators into the PSO algorithm to find an appropriate network architecture and connection weights. Fig.2 shows flowchart of the proposed algorithm. The major steps of the proposed algorithm are as follows:

1. Generate initial population. The hidden nodes are encoded as binary code string, 1 with connection and 0 without connection. The connection weights are encoded as float string, randomly generated within $[-1, 1]$.
2. Input training data and calculate the fitness of each particle according to Equation (4). Initialize individual best position $P_{best}(t)$ and the global best position $P_{gbest}(t)$.
3. Perform PSO operators. Compare individual current fitness and the fitness of its experienced best position. If current fitness is better, we set current position to be the best position. Compare individual current fitness and the fitness of the global best position. If current fitness is better, we set current position to be the global best position
4. Each individual updates its velocity according to Equation 3

$$v_{id}^{k+1} = \omega v_{id}^{k} + c_1\gamma_1(p_{id}^{k} - x_{id}^{k}) + c_2\gamma_2(p_{gd}^{k} - x_{id}^{k}) \tag{3}$$

   where v_{id}^{k+1} and x_{id}^{k} stand for the velocity and position of the ith particle of the kth iteration, respectively. p_{id}^{k} denotes the previously best position of particle i, p_{gd}^{k} denotes the global best position of the swarm. ω is the inertia weight, c_1 and c_2 are acceleration constants (the general value of c_1 and c_2 are in the interval [0 2]), γ_1 and γ_2 are random numbers in the range [0 1].

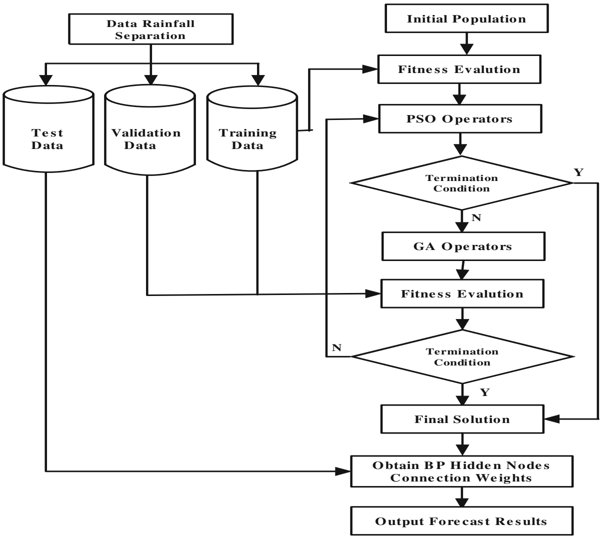


Fig. 2. Flowchart of algorithm

5. Each individual updates its position velocity according to Equations 4 and 5 for float string and binary code string, respectively.

$$x_{id}^{k+1} = x_{id}^{k} + v_{id}^{k+1} \tag{4}$$

$$x_{id}^{k+1} = \begin{cases} 1 & r < 1/(1 + exp(v_{id}^{k+1})) \\ 0 & else \end{cases} \tag{5}$$

where x_{id}^{k+1} stands for the position of the ith particle of the $k+1$th iteration, r denotes independently uniformly distributed random variable with range [0,1].

6. Judge termination. If an updated individual with new fitness cannot satisfy termination condition, go to step 7, otherwise the process output the final solution.
7. Perform GA process. The selection operator of genetic algorithm is implemented by using the roulette–wheel algorithm to determine which population members are chosen as parents that will create offspring for the next generation.
8. Use basic crossover and mutation operations to the binary code string, namely, if a hidden node is deleted (added) according to mutation operation, the corresponding control code is encoded 0 (1). The paper uses arithmetical crossover which can ensure the offspring are still in the constraint region and moreover the system is more stable and the variance of the best solution is smaller. The crossover of connection weights are operated with probability p_c at float string according to Equations 6 and 7

$$x_i^{t+1} = \alpha x_i^t + (1-\alpha)x_{i+1}^t \tag{6}$$

$$x_{i+1}^{t+1} = (1-\alpha)x_i^t + \alpha x_{i+1}^t \tag{7}$$

where x_i^{t+1} stands for the real values of the ith individual of the $(t+1)$th generation, x_i^t and x_{i+1}^t are a pair of individuals before crossover, x_i^{t+1} and x_{i+1}^{t+1} are a pair of individuals after crossover, α is taken as random value within [0, 1].

9. The mutation of connection weights are operated with probability p_m at float string according to Equation 10

$$x_i^{t+1} = x_i^t + \beta \tag{8}$$

where x_i^t stands for the real values of the ith individual of the tth generation, x_i^t is individual before mutation, x_i^{t+1} is individual after mutation, β is taken as random value within [0, 1].
10. Input training data and validation data, compute fitness for all the individuals by a fitness function.
11. Judge termination. Once the termination condition is met, output the final solution, Obtain the appropriate network architecture and connection weights, otherwise go to step 3. The maximum number of iterations is considered as the termination criterion.
12. Input testing data and Output forecasting results by the evolutionary NN.

3 Experimental Results and Discussion

The platform adopted to develop the PSOGA–NN approach is a PC with the following features: Intel Celeron M 1.86 GHz CPU, 1.5 GB RAM, a Windows XP operating system and the MATLAB development environment. Table 1 gives overview of PSOGA–NN parameter settings.

Table 1. The setting of parameter values for the proposed PSOGA–NN algorithm

The description of parameter	Value
The total number of training (epochs)	100
Population size	40
PSO algorithm part	
Inertia weight (ω)	2
Learning factor (c_1, c_2)	$c_1 = c_2 = 2$
The maximum velocity of each particle (V_{max})	3
GA algorithm part	
Crossover probability (p_c)	0.8
Mutation probability (p_m)	0.05

3.1 Study Area and Data

Real–time ground rainfall data have been obtained in monthly rainfall, which was collected from 10 stations of the Guilin Meteorology Administration rain gauge networks for the period from 1991 to 2009. Thus the training data set contains 252 data points, whose training set is 144 (1991–2002), validation set is 72 (2003-2008), and testing set is 36 (2007–2009).

3.2 Criteria for Evaluating Model Performance

Three different types of standard statistical performance evaluation criteria were employed to evaluate the performance of various models developed in this paper. These are average absolute relative error (AARE), root mean square error (RMSE), and the Pearson Relative Coefficient (PRC) which be found in many paper [17].

The input vector is represented by rainfall and runoff values for the preceding 6 monthly rainfall (i.e., $t-1, t-2, t-3, t-4, t-5, t-6$). Accordingly, the output vector represents the expected rainfall value for mothly $t(\hat{y_t})$. For the purpose of comparison by the same six input variables, we have also built other three rainfall forecasting models: multi–layer perceptron neural network model based on the back–propagation learning algorithm (BP–NN), pure genetic algorithm evolutionary neural network method(GA–NN) [18] proposed by Irani et al, and pure particle swarm optimization evolutionary neural network method (PSO–NN) [20] proposed by Chau et al. In this paper data, the best NN architecture is: 6–8–1 (6 input units, 8 hidden neurons, 1 output neuron), and the best NN parameters is chosen as a benchmark model for comparison by the trial–and–error method with the minimum testing root mean square error. Before training and testing all source data are normalized into the range between and 1, by using the maximum and minimum values of the variable over the whole data sets.

3.3 Analysis of the Results

Fig.3 shows the comparison between observed and predicted rainfall values at 36 validation samples by four different models model using the monthly rainfall data. From the Fig. 3, the output of the PSOGA–NN model, simulated with validation data, shows a good agreement with the target. We can find that the PSOGA–NN methods have better results than the methods only using BP–NN, GA–NN or PSO–NN in accuracy. Especially PSOGA–NN, it has the smallest error rates and standard deviation.

Table 2 illustrates the fitting, validation and testing accuracy and efficiency of the model in terms of various evaluation indices for training, validation and testing samples, respectively. From the table 2, we can generally see that learning ability of PSOGA–NN outperforms the other three models under the same network input. As a consequence, poor performance indices in terms of AARE, RMSE and PRC can be observed in BP–NN model than other three model. Table 2 also shows that the performance of PSOGA–NN is the best in case study for training samples, validation samples and testing samples. The more important factor to measure performance of a method is to check its forecasting ability of testing samples in order for actual rainfall application. Table 2 indicates that PSOGA–NN not only has the smallest AARE, but also the smallest RMSE. This means that PSOGA–NN is the most stable method. In addition, the PRC of PSOGA–NN is the maximum in all models. The results indicate that the deviations between observed and forecasting value are very small, are capable to capture the average change tendency of the mothly rainfall data.

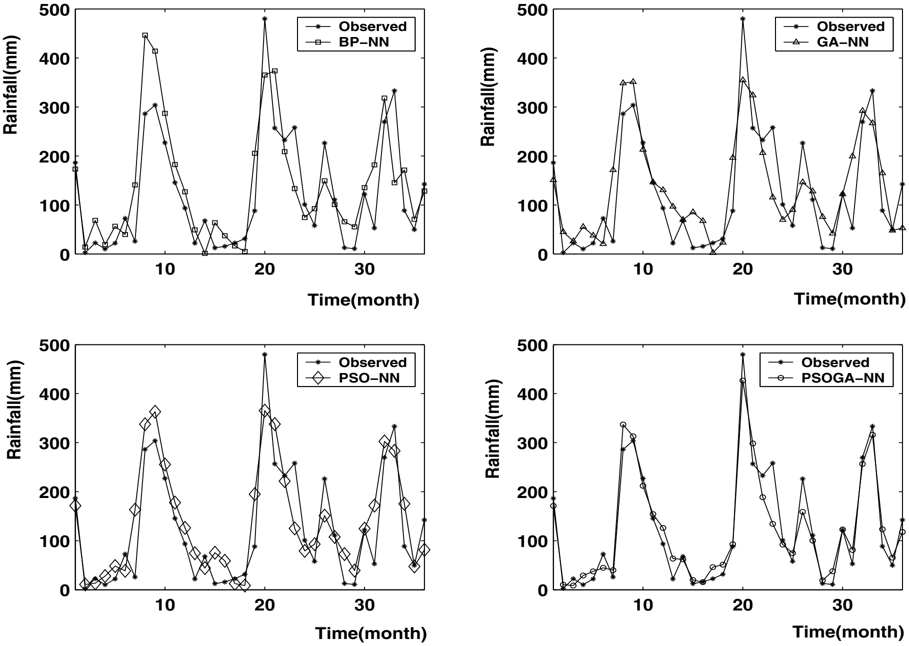


Fig. 3. Comparison between observed and predicted of model in testing samples

Table 2. Performance statistics of the five models for rainfall fitting and forecasting

Model	BP–NN	GA–NN	PSO–NN	PSOGA–NN
Index	**Training data (from 1991 to 2002)**			
AARE(%)	126.55(%)	86.64(%)	85.23(%)	49.25(%)
RMSE	80.9917	80.2205	70.2709	42.6571
PRC	0.9019	0.9028	0.9284	0.9736
	Validation data (from 2003 to 2006)			
AARE(%)	90.97(%)	110.18(%)	103.31(%)	38.83(%)
RMSE	59.1042	61.5616	58.8813	34.2263
PRC	0.8965	0.8871	0.9000	0.9629
	Testing data (from 2007 to 2009)			
AARE(%)	117.09	118.85	107.10	76.25
RMSE	85.96	67.98	68.93	67.92
PRC	0.7920	0.8662	0.8747	0.9636

From the graphs and table, we can generally see that the forecasting results are very promising in the rainfall forecasting under the research where either the measurement of fitting performance is goodness or where the forecasting performance is effectiveness. It also can be seen that there was consistency in the results obtained between the training and testing of these PSOGA–NN model. In this paper, the GA operators are integrated into the PSO algorithm to improve the NN performance. PSOGA–NN though integrating GA and PSO has been validated using three examples adopted from the literature. The experimental results indicate that PSOGA methods are always better than only using pure GA and pure PSO algorithm for NN evlutionary.

4 Conclusion

Accurate rainfall forecasting is crucial for a frequent unanticipated flash flood region to avoid life losing and economic loses. In this paper, the GA operators are integrated into the PSO algorithm to improve the evlutionary NN performance, avoiding problems of premature convergence, permutation and escaping from local optima. Experimental results with the Guilin monthly rainfall dataset suggest that the proposed strategy can improve the precidetion accuracy. The results demonstrate that the proposed three hybrid methods are able to provide better performance than GA and PSO for NN. It demonstrated that it increased the rainfall forecasting accuracy more than any other model employed in this study in terms of the same measurements. So the PSOGA–NN model can be used as an alternative tool for monthly rainfall forecasting to obtain greater forecasting accuracy and improve the prediction quality further in view of empirical results.

Acknowledgment. The authors would like to express their sincere thanks to the editor and anonymous reviewers comments and suggestions for the improvement of this paper. This work was supported in part by the Natural Science Foundation of China under Grant No.11161029, and in part by the Education Department of Guangxi under Grant No.201203YB186.

References

1. 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)
2. Wu, J., Jin, L.: Study on the Meteorological Prediction Model Using the Learning Algorithm of Neural Networks Ensemble Based on PSO agorithm. Journal of Tropical Meteorology 15(1), 83–88 (2009)
3. Gwangseob, K., Ana, P.B.: Quantitative Flood Forecasting Using Multisensor Data and Neural Networks. Journal of Hydrology 246, 45–62 (2001)
4. 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 III. LNCS, vol. 5553, pp. 49–58. Springer, Heidelberg (2009)
5. Wu, J.: An Effective Hybrid Semi-Parametric Regression Strategy for Rainfall Forecasting Combining Linear and Nonlinear Regression. International Journal of Applied Evolutionary Computation 2(4), 50–65 (2011)

6. Wu, J.: A Semiparametric 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)
7. Sedki, A., Ouazar, D., Mazoudi, E.E.: Evolving Neural Network Using Real Coded Genetic Algorithm for Daily Rainfall–runoff Forecasting. Expert Systems with Applications 36, 4523–4527 (2009)
8. Malinak, P., Jaksa, R.: Simultaneous Gradient and Evolutionary Neural Network Weights Adaptation Methods. In: IEEE Congresson Evolutionary Computation (CEC), September 25-28 (2007)
9. 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)
10. Kiranyaz, S., Ince, T., Yildirim, A., Gabbouja, M.: Evolutionary Artificial Neural Networks by Multi–dimensional Particle Swarm Optimization. Neural Networks 22, 1448–1462 (2009)
11. Wu, J., Jin, L., Liu, M.: Modeling Meteorological Prediction Using Particle Swarm Optimization and Neural Network Ensemble. In: Wang, J., Yi, Z., Żurada, J.M., Lu, B.-L., Yin, H. (eds.) ISNN 2006. LNCS, vol. 3973, pp. 1202–1209. Springer, Heidelberg (2006)
12. Oysu, C., Bingul, Z.: Application of Heuristic and Hybrid-GASA Algorithms to Tool–path Optimization Problem for Minimizing Airtime during Machining. Engineering Applications of Artificial Intelligence 22, 389–396 (2009)
13. Kennedy, J., Mendes, R.: Neighborhood Topologies in Fully-informed and Bestof-neighborhood Particles Swarms. In: Proceedings of the IEEE International Workshop on Soft Computing in Industrial Applications, pp. 45–50 (2003)
14. Kumanan, S., Jose, G.J., Raja, K.: Multi-project Scheduling Using an Heuristic and a Genetic Algorithm. Journal of Advanced Manufacturing Technology 31, 360–366 (2006)
15. Chen, P.H., Shahandashti, S.M.: Hybrid of Genetic Algorithm and Simulated Annealing for Multiple Project Scheduling with Multiple Resource Constraints. Automation in Construction 18, 434–443 (2009)
16. Babaoglu, I., Findik, O., Ülker, E.: A Comparison of Feature Selection Models Utilizing Binary Particle Swarm Optimization and Genetic Algorithm in Determining Coronary Artery Disease Using Support Vector Machine. Expert Systems with Applications 37, 3177–3183 (2010)
17. Wang, K., Yang, J., Shi, G., Wang, Q.: An Expanded Training Set Based Validation Method to Avoid Over Fitting for Neural Network Classifier. In: Fourth International Conference on Natural Computation, vol. 3, pp. 83–87 (2008)
18. Irani, R., Nasimi, R.: Evolving Neural Network Using Real Coded Genetic Algorithm for Permeability Estimation of The Reservoir. Expert Systems with Applications 38, 9862–9866 (2011)
19. Lin, Y.C., Zhang, J., Zhong, J.: Application of Neural Networks to Predict The Elevated Temperature Behavior of a Low Alloy Steel. Computational Material Science 43(4), 752–758 (2008)
20. Chau, K.W.: Particle Swarm Optimization Training Algorithm for ANNs in Stage Prediction of Shing Mun River. Journal of Hydrology 329, 363–367 (2006)

Forecasting the TAIEX Based on Fuzzy Time Series, PSO Techniques and Support Vector Machines

Shyi-Ming Chen and Pei-Yuan Kao

Department of Computer Science and Information Engineering,
National Taiwan University of Science and Technology,
Taipei, Taiwan, R.O.C.

Abstract. This paper presents a new method for forecasting the TAIEX based on fuzzy time series, particle swarm optimization techniques and support vector machines. The proposed method to forecast the TAIEX is based on slope of one-day variations of the TAIEX and the slope of two-days average variations of the TAIEX. The particle swarm optimization techniques are used to get optimal intervals in the universe of discourse. The support vector machine is used to classify the training data set. The experimental results show that the proposed method outperforms the existing methods for forecasting the TAIEX.

Keywords: Fuzzy Sets, Fuzzy Time Series, Support Vector Machines, TAIEX.

1 Introduction

In [15], [16] and [17], Song and Chissom proposed the concepts of fuzzy time series. In [16] and [17], Song and Chissom presented the time-invariant fuzzy time series model and the time-variant fuzzy time series model to forecast the enrollments of the University of Alabama. In recent years, some fuzzy forecasting methods based on fuzzy time series have been presented [2], [4]-[8], [11]-[13], [18]-[19].

In this paper, we present a new method for forecasting the Taiwan Stock Exchange Capitalization Weighted Stock Index (TAIEX) based on fuzzy time series, particle warm optimization techniques [10] and support vector machines [20]. The proposed method to forecast the TAIEX is based on the slope of one-day variations of the TAIEX and the slope of two-days average variations of the TAIEX. Because the slope of two-days average variations of the TAIEX is smoother than the slope of one-day variations of the TAIEX, we choose it to define the universe of discourse. The particle swarm optimization techniques are used to get optimal intervals in the universe of discourse. The support vector machine is used to classify the training data set. The first feature and the second feature of the support vector machine are the slope of one-day variations and the slope of two-days variations of the TAIEX, respectively. The experimental results show that the proposed method outperforms the existing methods for forecasting the TAIEX.

A. Selamat et al. (Eds.): ACIIDS 2013, Part I, LNAI 7802, pp. 89–98, 2013.

2 Fuzzy Time Series

In this section, we briefly review some basic concepts of fuzzy time series from [3], [15]-[17], where the values of fuzzy time series are represented by fuzzy sets [21]. Let U be the universe of discourse, where $U = \{ u_1, u_2, ..., u_n \}$. A fuzzy set A_i in the universe of discourse U can be represented by

$$A_i = f_{A_i}(u_1)/u_1 + f_{A_i}(u_2)/u_2 + \cdots + f_{A_i}(u_n)/u_n, \tag{1}$$

where f_{A_i} is the membership function of the fuzzy set A_i, $f_{A_i}(u_j)$ denotes the degree of membership of u_j belonging to the fuzzy set A_i, $f_{A_i}(u_j) \in [0, 1]$ and $1 \leq j \leq n$.

Let $Y(t)$ $(t = ..., 0, 1, 2, ...)$, a subset of R, be the universe of discourse in which fuzzy sets $f_i(t)$ $(i = 1, 2, ...)$ are defined and let $F(t)$ be a collection of $f_i(t)$ $(i = 1, 2, ...)$. Then, $F(t)$ is called a fuzzy time series on $Y(t)$ $(t = ..., 0, 1, 2, ...)$. If there exists a fuzzy logical relationship $R(t-1, t)$, such that $F(t) = F(t-1) \circ R(t, t-1)$, where both $F(t)$ and $F(t-1)$ are fuzzy sets and the symbol "$\circ$" is the max-min composition operator, then $F(t)$ is called derived by $F(t-1)$, denoted by a fuzzy logical relationship shown as follows:

$$F(t-1) \rightarrow F(t).$$

If $F(t-1) = A_i$ and $F(t) = A_j$, where A_i and A_j are fuzzy sets, then the fuzzy logical relationship between $F(t-1)$ and $F(t)$ can be represented by

$$A_i \rightarrow A_j,$$

where A_i and A_j are called the current state and the next state of the fuzzy logical relationship, respectively.

3 Particle Swarm Optimization

In [10], Kennedy and Eberhart developed an optimization algorithm, named particle swarm optimization (PSO), which was inspired by the social behavior of bird flocking or fish schooling. In PSO, a set of particles consist of particle swarms, where a particle denotes a potential solution. The position vector and the velocity vector of the ith particle in the n-dimensional search space can be represented as $X_i = [x_{i,1}, x_{i,2}, \cdots, x_{i,n}]$ and $V_i = [v_{i,1}, v_{i,2}, \cdots, v_{i,n}]$, respectively. Each particle has the personal best position P_{best} which records the position of the best corresponding objective value of the particle. The global position P_{gbest} denotes the position of the best particle in the swarm. The new velocity and current position of each particle is updated as follows:

$$V_{i,t} = \omega \times V_{i,t-1} + C_1 \times r_1 \times (P_{best,i} - X_{i,t-1}) + C_2 \times r_2 \times (P_{gbest} - X_{i,t-1}) \tag{2}$$

$$X_{i,t} = X_{i,t-1} + V_{i,t}, \tag{3}$$

where $V_{i,t}$ denotes the velocity of *i*th particle at the *t*th iteration and $V_{i,t}$ is limited to a pre-defined range [-V_{MAX}, V_{MAX}]. The symbol ω denotes the inertial weight coefficient, and C_1 and C_2 are the acceleration coefficients, r_1 and r_2 are two independent random numbers uniformly distributed in the range of [0, 1]. The procedure of a PSO algorithm is shown as follows:

Step 1: Initialize all particles with their respective position and velocity.

Step 2: Update the personal best position $P_{best,i}$ of each *i*th particle and find the best particle P_{gbest} and its position P_{gbest} in the swarm.

Step 3: Update each particle's velocity and position based on Eqs. (2) and (3), respectively.

Step 4: Repeatedly perform Step 2 and Step 3 until the stop condition is reached.

4 An New Method for Forecasting the TAIEX Based on Fuzzy Time Series, Particle Swarm Optimization Techniques and Support Vector Machines

In this section, we present a new method to forecast the TAIEX [22] based on fuzzy time series, particle swarm optimization techniques and support vector machines. The proposed method is now presented as follows:

Step 1: Let $V1_t$ denote the variation between day *t* and day *t-1* and let $V2_t$ denote the average variation between day *t* and day *t-2*, shown as follows:

$$V1_t = TAIEX_t - TAIEX_{t-1}, \tag{4}$$

$$V2_t = (TAIEX_t - TAIEX_{t-2})/2, \tag{5}$$

where $TAIEX_{t-2}$, $TAIEX_{t-1}$ and $TAIEX_t$ donote the TAIEX on day *t-2*, day *t-1* and day *t*, respectively. Calculate the slope of variation between day *t* and day *t-1* and calculate the slope of the average variation between day *t* and day *t-2*. Let *S1* and *S2* denote the slope of one-day variation of the TAIEX and the slope of two-days average variation of the TAIEX on trading day *t*, respectively, shown as follows:

$$S1 = V1_t - V1_{t-1}, \tag{6}$$

$$S2 = V2_t - V2_{t-1}, \tag{7}$$

Step 2: Define the universe of discourse *U*, $U = [S2_{min} - p, S2_{max} + q]$, where $S2_{min}$ and $S2_{max}$ are the minimum value and the maximum value of the slope of two-days average variations *S2* of the TAIEX of historical training data, respectively; *p* and *q* are two proper positive real values.

Step 3: Use particle swarm optimization (PSO) techniques to get the optimal split point vector $S = [p_1, p_2, \ldots, p_n]$, such that the universe of discourse *U* is divided into *n*+1 intervals, i.e., $[0, p_1)$, $[p_1, p_2)$, ..., $[p_{n-1}, p_n]$, where the sub-steps are shown as follows:

Step 3.1: Generate *m* particles with dimension *n*, where *n* denotes number of split points. The position of the *i*th particle is denoted by X_i consisting of elements $x_{i,1}, x_{i,2}, \cdots,$ and $x_{i,n}$, where $x_{i,j} \in [S2_{min}, S2_{max}]$, $1 \le j \le n$ and $x_{i,1} < x_{i,2} < \cdots < x_{i,n}$, as shown in Figure 1. The velocity of the *i*th particle is denoted by V_{i} consisting of elements $v_{i,1}, v_{i,2}, \cdots,$ and $v_{i,n}$ where $v_{i,j} \in [-(S2_{max} - S2_{min}), S2_{max} - S2_{min}]$ and $1 \le j \le n$, as shown in Fig. 2.

$x_{i,1}$	$x_{i,2}$	$\cdots$	$x_{i,n}$

Fig. 1. Graphical representation of the position vector *Xi* of the *i*th particle

$v_{i,1}$	$v_{i,2}$	$\cdots$	$v_{i,n}$

Fig. 2. Graphical representation of the velocity vector *Vi* of the *i*th particle

At the beginning, the system generates *m* particles with the randomly generated initial position vector X_i and the randomly generated initial velocity vector V_i for the *i*th particle, where $1 \le i \le m$. The personal best position vector $P_{best,i} = [b_{i,1}, b_{i,2}, \cdots, b_{i,n}]$ denotes the best position (i.e., the position that gives the minimum objective value) of the *i*th particle found so far. Therefore, the personal best position vector $P_{best,i}$ of the *i*th particle is the same as its initial position vector at the beginning.

Step 3.2: Set the number of iteration *iter* to zero (i.e., let *iter* = 0). Then, calculate the objective value of each particle, update the personal best position of each particle, choose the best particle among all particles and update the velocity and position of each particle, and then increase the number of iterations *iter* by 1 for each iteration, until the number of iterations *iter* reaches a predefined number of iterations. The detailed sub-steps of this step are shown as follow:

Step 3.2.1: Calculate the objective value of each particle. Let $diff_j$ denotes the sum of the distance between data in the *j*th interval, where $1 \le j \le n+1$. For example, assume that there are three data *x*, *y* and *z* in the *j*th interval, then

$$diff_j = |x - y| + |x - z| + |y - z|. \tag{8}$$

Calculate the average difference $average\_diff_i$ of the *i*th particle, shown as follows:

$$average\_diff_i = \frac{diff_{i,1} + diff_{i,2} + \cdots + diff_{i,n+1}}{n+1}. \tag{9}$$

Let the average difference $average\_diff_i$ be the objective value of the *i*th particle.

Step 3.2.2: If the objective value of the *i*th particle with the position vector $X_i = [x_{i,1}, x_{i,2}, \cdots x_{i,n}]$ at the current iteration *iter* is smaller than the objective value of $P_{best,i}$, then assign the values of the elements in the position vector X_i at the current iteration to the elements in the personal best position vector $P_{best,i} = [b_{i,1}, b_{i,2}, \cdots, b_{i,n}]$, shown as follows:

$$\begin{gathered} b_{i,1} = x_{i,1}, \\ b_{i,2} = x_{i,2}, \\ \vdots \\ b_{i,n} = x_{i,n}. \end{gathered}$$

Step 3.2.3: Update P_{gbest}. If the objective value of $P_{best,i}$ is smaller than the objective value of P_{gbest}, then let $P_{gbest} = P_{best,i}$, where $1 \le i \le m$ and P_{gbest} denote the position vector of the best particle *gbest.*

Step 3.3: Update the velocity vector V_i and the position vector X_i of the *i*th particle based on Eqs. (2) and (3), where $1 \le i \le m$. In this paper, we let $C_1 = 2$, $C_2 = 2$ and $\omega = 1$.

Step 3.4: If the number of iterations *iter* is smaller than a predefined number of iterations, then increase *iter* by 1 and go to **Step 3.1**. Otherwise, let the position vector P_{gbest} of the best particle *gbest* obtained in **Step 3.2.3** be the optimal split point vector $S = [p_1, p_2, \ldots, p_n]$.

Step 4: Define the linguistic terms A_1, A_2, …, and A_n represented by fuzzy sets based on the *n* intervals u_1, u_2, …, and u_n obtained in **Step 3**, shown as follows:

$$\begin{aligned} A_1 &= 1/u_1 + 0.5/u_2 + 0/u_3 + \cdots + 0/u_{n-2} + 0/u_{n-1} + 0/u_n, \\ A_2 &= 0.5/u_1 + 1/u_2 + 0.5/u_3 + \cdots + 0/u_{n-2} + 0/u_{n-1} + 0/u_n, \\ &\vdots \\ A_{n-1} &= 0/u_1 + 0/u_2 + 0/u_3 + \cdots + 0.5/u_{n-2} + 1/u_{n-1} + 0.5/u_n, \\ A_n &= 0/u_1 + 0/u_2 + 0/u_3 + \cdots + 0/u_{n-2} + 0.5/u_{n-1} + 1/u_n. \end{aligned}$$

Step 5: Fuzzify each historical training datum of the two-days average variations (i.e., *S2*) of the TAIEX into a fuzzy set defined in **Step 4**. If the historical training datum of the two-days average variation (i.e., $S2_t$) of the TAIEX on trading day *t* belongs to u_i and the maximum membership value of the fuzzy set A_i occurs at interval u_i, where $1 \le i \le n$, then the historical training datum of the two-days average variation (i.e., *S2*) of the TAIEX on trading day *t* is fuzzified into A_i.

Step 6: Fuzzify each historical testing datum of two-days average variations (i.e., *S2*) of the TAIEX into a fuzzy set defined in **Step 4.**

Step 7: Construct fuzzy logical relationships from the fuzzified historical training data of the two-days average variation of the TAIEX obtained in **Step 5**. For example, if the sequence of the fuzzified historical training data is: A_1, A_2, A_3, A_1, A_3, A_1, then the following fuzzy logical relationships are generated:

$$
\begin{aligned}
&A_1 \rightarrow A_2,\\
&A_2 \rightarrow A_3,\\
&A_3 \rightarrow A_1,\\
&A_1 \rightarrow A_3,\\
&A_3 \rightarrow A_1.
\end{aligned}
$$

Step 8: Construct fuzzy logical relationship groups based on the fuzzy logical relationships obtained in **Step 7**, where if the fuzzy logical relationships have the same antecedent, then these fuzzy logical relationships can be grouped into a fuzzy logical relationship group. For example, because the following fuzzy logical relationships have the same antecedent "A_1":

$$A_1 \rightarrow A_2,$$

$$A_1 \rightarrow A_3,$$

then they can be grouped into a fuzzy logical relationship group, shown as follows:

$$A_1 \rightarrow A_2, A_3.$$

Step 9: If the fuzzified historical testing datum on trading day t matches the antecedents of more than one fuzzy logical relationships (i.e., it matches the antecedent of a fuzzy logical relationship group), then we use the support vector machine to forecast the value of "fuzzified $S2$" of the next trading day $t+1$, where the first feature and the second feature of the support vector machine are "the slope of one-day variations" and "the slope of two-days average variations" of the TAIEX on trading day t, respectively, and the target of the support vector machine is the fuzzy set of the slope of "two-days average variations" (i.e., "fuzzified $S2$") of the TAIEX of the trading day $t+1$. For example, assume that the training data is shown as follows:

Day	$S1$	$S2$	Fuzzified $S2$
t	100	50	A_1
$t+1$	300	75	A_2
$t+2$	250	150	A_3
$t+3$	50	25	A_1
$t+4$	200	125	A_3
$t+5$	110	45	A_1

Then, the following fuzzy logical relationships are generated:

$$A_1 \rightarrow A_2,$$

$$A_2 \rightarrow A_3,$$

$A_3 \rightarrow A_1$,

$A_1 \rightarrow A_3$,

$A_3 \rightarrow A_1$.

Assume that the testing datum is shown as follows:

Day	*S1*	*S2*	Fuzzified *S2*
t +6	unknown	unknown	?

where the symbol "?" denotes the fuzzy set in which we want to predict for the trading day *t+6*. Because the fuzzified *S2* on trading day *t+5* is A_1, we can see that the fuzzy set A_1 matches the antecedent of the following generated fuzzy logical relationships:

$A_1 \rightarrow A_2$,
$A_1 \rightarrow A_3$,

where these two fuzzy logical relationships form the fuzzy logical relationship group, shown as follows:

$A_1 \rightarrow A_2, A_3$.

Then, we convert the forecasting problem into the classification problem. Because the instances in the training data set whose fuzzified *S2* is A_1 are shown as follows:

Day	*S1*	*S2*	Fuzzified *S2*
t	100	50	A_1
t+3	50	25	A_1

we let the fuzzified *S2* of the above instances on trading day *t* be equal to the fuzzified *S2* of the next trading day *t+1*, shown as follows:

Day	*S1*	*S2*	Fuzzified *S2*
t	100	50	A_2
t+3	50	25	A_3

Assume that the testing datum is shown as follows:

Day	*S1*	*S2*	Fuzzified *S2*
t +6	110	45	?

where the symbol "?" is the fuzzy set in which we want to predict. It should be noted that there are two parameters which can influence the classification result of a support vector machine, where the one is called the "cost parameter" and the other one is called the "gamma parameter". We tune the above two parameters by using the grid search method [1] and the cross-validation method [1]. Let the classification result by using the support vector machine be the testing datum's "Fuzzified *S2*" in which we want to predict. Otherwise, if the fuzzified *S2* of the testing datum on trading day *t* does not match the antecedent of any fuzzy logical relationships (i.e., it does not match the antecedent of any fuzzy logical relationship groups), then we let the predicted fuzzified *S2* of the testing datum on trading day *t+1* be equal to the fuzzified *S2* of the testing datum on trading day *t*.

Step 10: If the forecasted fuzzified *S2* on trading day *t+1* obtained in **Step 9** is A_k and the maximum membership value of A_k occurs at interval u_k, then let the defuzzified value *D* of the fuzzified *S2* be equal to the midpoint of interval u_k. For example, assume that the fuzzified *S2* of trading day *t+1* obtained in **Step 9** is A_1, where the maximum membership value of the fuzzy set A_1 occurs at interval u_1, u_1 = [0, 50] and the midpoint of the interval u_1 is 25, then the defuzzified value *D* of the fuzzified *S2* on trading day *t+1* is 25.

Step 11: Use the defuzzified value *D* of the fuzzified *S2* obtained in **Step 10** to forecast the TAIEX of trading day *t+1* (i.e., $TAIEX_{t+1}$), where

$$TAIEX_{t+1} = 2 \times (D + V2_t) - V1_t + TAIEX_t \tag{10}$$

and $TAIEX_t$ denotes the TAIEX on trading day *t*.

5 Experimental Results

In this section, we apply the proposed method to forecast the TAIEX from 1999 to 2004. In order to compare the experimental results of the proposed method with the ones of the methods presented in [3], [6], [7], [8], [18] and [19], we also adopt a 10 month/2 month split for training/testing, i.e., for each year, the data from January to October are used as the training data set, and the data from November to December are used as the testing data set. We use the Root Mean Squared Error (RMSE) to evaluate the performance of the proposed method, defined as follows:

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n} forcasted\ value_i - actual\ value_i}{n}}\ , \tag{11}$$

where *n* denotes the number of trading days needed to be forecasted. The system executed the proposed method 100 times to get the average RMSE. Table 1 shows the RMSEs and the average RMSE of the proposed method for forecasting the TAIEX from 1999 to 2004. Table 2 shows the RMSEs and the average RMSE of the proposed method for forecasting the TAIEX from 1990 to 1999. From Table 1 and Table 2, we can see that the average RMSE of the proposed method is smaller than that of the existing methods [3], [6], [7], [8], [18] and [19]. In other words, the proposed method outperforms the methods presented in [3], [6], [7], [8], [18] and [19] for forecasting the TAIEX.

Table 1. The RMSEs and the average RMSE of the proposed method for forecasting the TAIEX from 1999 to 2004

RMSEs Year / Method	1999	2000	2001	2002	2003	2004	Average RMSE
The Proposed Method	87.63	125.34	114.57	76.86	54.29	58.17	86.14

Table 2. The RMSEs and the average RMSE of the proposed method for forecasting the TAIEX from 1990 to 1999

RMSEs Year / Method	1990	1991	1992	1993	1994	1995	1996	1997	1998	1999	Average RMSE
The Proposed Method	156.47	56.50	36.45	126.45	105.52	62.57	51.50	125.33	104.12	87.63	91.25

6 Conclusions

We have presented a new method for forecasting the TAIEX based on fuzzy time series, particle swarm optimization techniques and support vector machines. The proposed method forecast the TAIEX based on slope of one-day variations of the TAIEX and the slope of two-days average variations of the TAIEX. The particle swarm optimization techniques are used to get optimal intervals in the universe of discourse. The support vector machine is used to classify the training data set. From Table 1 and Table 2, we can see that the average RMSE of the proposed method is smaller than that of the existing methods [3], [6], [7], [8], [18] and [19].

Acknowledgements. The authors work like to thank Professor Tian-Wei Sheu for his help during this work. This work was supported by the National Science Council Taiwan, Republic of China, under Grant NSC 101-2221-E- 011-171-MY2.

References

1. Chang, C.C., Lin, C.J.: LIBSVM: A Library for Support Vector Machines (2001), `http://www.csie.ntu.edu.tw/~cjlin/libsvm`
2. Chen, C.D., Chen, S.M.: A New Method to Forecast the TAIEX Based on Fuzzy Time Series. In: Proc. 2009 IEEE Int. Conf. Syst., Man, Cyber., San Antonio, Texas, pp. 3550–3555 (2009)
3. Chen, S.M.: Forecasting Enrollments Based on Fuzzy Time Series. Fuzzy Sets and Systems 81(3), 311–319 (1996)
4. Chen, S.M.: Forecasting Enrollments Based on High-Order Fuzzy Time Series. Cybernetics and Systems 33(1), 1–16 (2002)

5. Chen, S.M., Chang, Y.C.: Multi-Variable Fuzzy Forecasting Based on Fuzzy Clustering and Fuzzy Interpolation Techniques. Information Sciences 180(24), 4772–4783 (2010)
6. Chen, S.M., Chen, C.D.: TAIEX Forecasting Based on Fuzzy Time Series and Fuzzy Variation Groups. IEEE Transactions on Fuzzy Systems 19(1), 1–12 (2011)
7. Chen, S.M., Manalu, G.M.T., Shih, S.C., Sheu, T.W., Liu, H.C.: A New Method for Fuzzy Forecasting Based on Two-Factors High-Order Fuzzy-Trend Logical Relationship Groups and Particle Swarm Optimization Techniques. In: Proceedings of the 2011 IEEE International Conference on Systems, Man, and Cybernetics, Anchorage, Alaska, pp. 2301–2306 (2011)
8. Huarng, K.H., Yu, T.H.K., Hsu, Y.W.: A Multivariate Heuristic Model for Fuzzy Time-Series Forecasting. IEEE Transactions on Systems, Man, and Cybernetics- Part-B: Cybernetics 37(4), 836–846 (2007)
9. Hwang, J.R., Chen, S.M., Lee, C.H.: Handling Forecasting Problems Using Fuzzy Time Series. Fuzzy Sets and Systems 100(2), 217–228 (1998)
10. Kennedy, J., Eberhart, R.: Particle Swarm Optimization. In: Proceedings of the 1995 IEEE International Conference on Neural Networks, Perth, Australia, vol. 4, pp. 1942–1948 (1995)
11. Kuo, I.H., Horng, S.J., Chen, Y.H., Run, R.S., Kao, T.W., Chen, R.J., Lai, J.L., Lin, T.L.: Forecasting TAIFEX Based on Fuzzy Time Series and Particle Swarm Optimization. Expert Systems with Applications 37(2), 1494–1502 (2010)
12. Kuo, I.H., Horng, S.J., Kao, T.W., Lin, T.L., Lee, C.L., Pan, Y.: An Improved Method for Forecasting Enrollments Based on Fuzzy Time Series and Particle Swarm Optimization. Expert Systems with Applications 36(3), 6108–6117 (2009)
13. Lee, L., Wang, L.H., Chen, S.M., Leu, Y.H.: Handling Forecasting Problems Based on Two-Factors High-Order Fuzzy Time Series. IEEE Transactions on Fuzzy Systems 14(3), 468–477 (2006)
14. Sullivan, J., Woodall, W.H.: A Comparison of Fuzzy Forecasting and Markov Modeling. Fuzzy Sets and Systems 64(3), 279–293 (1994)
15. Song, Q., Chissom, B.S.: Fuzzy Time Series and Its Model. Fuzzy Sets and Systems 54(3), 269–277 (1993)
16. Song, Q., Chissom, B.S.: Forecasting Enrollments with Fuzzy Time Series-Part I. Fuzzy Sets and Systems 54(1), 1–9 (1993)
17. Song, Q., Chissom, B.S.: Forecasting Enrollments with Fuzzy Time Series-Part II. Fuzzy Sets and Systems 62(1), 1–8 (1994)
18. Yu, H.K.: Weighted Fuzzy Time-Series Models for TAIEX Forecasting. Physica A 349(3-4), 609–624 (2004)
19. Yu, T.H.K., Huarng, K.H.: A Bivariate Fuzzy Time Series Model to Forecast the TAIEX. Expert Systems with Applications 34(4), 2945–2952 (2008)
20. Vapnik, V.: Statistical Learning Theory. Wiley (1998)
21. Zadeh, L.A.: Fuzzy Sets. Information and Control 8, 338–353 (1965)
22. TAIEX Web Site,
`http://www.twse.com.tw/en/products/indices/tsec/taiex.php`

On the Design of Neighboring Fuzzy Median Filter for Removal of Impulse Noises

Chung-Ming Own[1,*] and Chi-Shu Huang[2]

[1] Department of Computer and Communication Engineering, St. John's University
[2] Department of Computer Science and Information Engineering,
National Chung Cheng University
cmown@mail.sju.edu.tw

Abstract. The digital images are easily affected by the noises; hence the image filters are often regarded as preprocessing of image processing system. If the image has serious damage or high-noise, the traditional image filters are usually unable to handle well. The application of median filter has been investigated. As an advanced method compared with standard median filtering, the adaptive median filter performs spatial processing to preserve detail and smooth non-impulsive noise. In this paper, a novel filter method, a neighboring selection method based on the fuzzy median filter, is proposed to improve the existing filter so that more image details can be preserved while effectively suppressing impulse noise. The proposed filter mechanism is composed of a new efficient noise eliminator based on the ideal of image rotation and LVQ network. Extensive simulation results demonstrate that our scheme performs significantly improve the default filter.

Keywords: Fuzzy Median Filter, LVQ network, Impulse Noise.

1 Introduction

Digital images are often corrupted by impulse noise when transmitted over a noisy channel or taken with a camera whose sensor is faulty 1. In the development of signal processing on removing impulse noises, median filter, one of the most popular nonlinear filters, has been extensively used for the removal of impulse noise due to its outstanding computational efficiency 2. However, the median filter tends to blur fine details and lines in many cases. Hence, modified forms of median based filters which still retain the rank order structure have been improved to overcome the deficiencies.

Among those are weighted median filter, which is an extension of the median filter, gives more weight to some values within the window 3, center-weighted median filters, which give the current pixel a large weight and the final output is chosen between the median and the current pixel value 4, furthermore, detail-preserving median filters 5, and rank-ordered mean filter 6, which excludes the current pixel itself from the operation window. Moreover, in the conventional conditional median filter, a threshold is set to separate the difference between the input signal and the median

A. Selamat et al. (Eds.): ACIIDS 2013, Part I, LNAI 7802, pp. 99–107, 2013.

value. However, this filter is still difficult to design, because the degree of impulse noises and the quantitative information are not easy to obtain.

Accordingly, about the weighted median filter, setting and analyzing the weights is difficult for actual signal processing. In 1996, Arakawa et al. presented a novel median-type filter controlled by fuzzy rules tried to overcome this deficiencies 7. This filter is constructed as a weighted sum of the input signal and the output of the median filter, and the weighting coefficients are set based on fuzzy rules concerning the state of the input signal sequence. The learning process is realized by training over a reference image to obtain the optimal weight. Therefore, one primary difficulty with this approach is the computational complexity associated with determining the weighting coefficients. Arakawa et al. restrict all region boundaries to the parallel to the coordinate axes to simply the complexity. In [8], the authors proposed an adaptive fuzzy switching filter to adopt a fuzzy logic approach for the enhancement of images corrupted by impulse noise. They developed the maximum-minimum exclusive median method to estimate the current pixel. Besides, the authors proposed another fuzzy based median filter to achieve improved filtering performance in terms of effectiveness in removing salt-and-pepper noise while preserving image details [9].

In this paper, the author proposes a novel method that improves the performance of the modified median filter by analyzing the spatial location of the weighting coefficients. Inspecting with different view of the weighting coefficients, the system can extend the inference result and obtain the more reliable result unnoticed.

The outline of this paper is as follows. Section 2 briefly reviews basic concepts of fuzzy median filter. Section 3 states the algorithm of our proposed intelligent mechanism. Section 4 presents empirical analyses of image restoration. Section 5 concludes this study.

2 Noise Model

Accordingly, the impulse noise model and general definition are reviewed briefly. In a variety of impulse noise models for images, corrupted pixels are often replaced with values equal to or near the maximum or minimum of the allowable dynamic rage. In the experiments, the author considers a general noise model in which a noisy pixel can take on arbitrary values in the dynamic rage according to some underlying probability distribution. Assume an image is treated as a two-dimensional matrix. Let $v(i,j)$ and $x(i,j)$ represent the luminance values of the original and the noisy image at location of pixel (i,j), respectively. Then, for an impulse noise model with error probability p_e is

$$x(i,j)=\begin{cases} v(i,j), & \text{with probability } 1 - p_e \\ p(i,j), & \text{with probability } p_e \end{cases} \tag{1}$$

where $p(i,j)$ is an identically distributed, independent random process with an arbitrary underlying probability density function.

Accordingly, define the real-valued 2-D sequence $\mathbf{w}(i,j)\in \Re^9$ as { $x(i-K,j-L)$, $x(i-K+1,j-L+1)$, …, $x(i,j)$, …, $x(i+K,j+L)$ } with a window size $(2K+1)\times(2L+1)$. The center signal of the window is $x(i,j)$. Then, the usual median filter puts out the median value, denoted as $m(i,j)$ in the $\mathbf{w}(i,j)$.

2.1 Fuzzy Noise Model

According to the Arakawa et al.'s proposition, the median filter based on fuzzy rules can be represented as follow,

$$y(i,j)=m(i,j)+\mu[\mathbf{w}(i,j)](x(i,j)-m(i,j)), \tag{1}$$

where $\mu[\mathbf{w}(i,j)]$ denotes the membership function indicating to what extent an impulse noise is considered not to be located at the pixel (i,j). $\mu[\mathbf{w}(i,j)]$ is determined by the state of the input signals in the filter window. That is, $\mu[\mathbf{w}(i,j)]=0$ indicates that an impulse noise is considered to be located at the pixel (i,j). Hence, the output of the filter is set equal to the median value of the input signals. Otherwise, $\mu[\mathbf{w}(i,j)]$ $=1$ indicates that an impulse noise is considered not to be located at the pixel (i,j). In this case, the output of the filter is set equal to the original input $x(i,j)$. Furthermore, $\mu[\mathbf{w}(i,j)]$ also takes a continue value from 0 to 1 to cope with the uncertain case which is difficult to judge whether an impulse noise exists or not.

Accordingly, because the membership function $\mu[\mathbf{w}(i,j)]$ can be set by the local characteristics of the input signals, the amplitudes of most noises can be used to indicate the degree of $\mu[\mathbf{w}(i,j)]$. Hence, the following rules are adopted as follow,

Rule 1: IF $u(i,j)$ is small and $v(i,j)$ is small, THEN $\mu[\mathbf{w}(i,j)]$ is large,
Rule 2: IF $u(i,j)$ is small and $v(i,j)$ is large, THEN $\mu[\mathbf{w}(i,j)]$ is small,
Rule 3: IF $u(i,j)$ is large and $v(i,j)$ is small, THEN $\mu[\mathbf{w}(i,j)]$ is small,
Rule 4: IF $u(i,j)$ is large and $v(i,j)$ is large, THEN $\mu[\mathbf{w}(i,j)]$ is very small,

where $u(i,j)$ is denoted as the absolute difference between the input $x(i,j)$ and the median value $m(i,j)$, that is $u(i,j)= |x(i,j) - m(i,j)|$. The indicator $u(i,j)$is simple and can remove prominent impulse noises. $v(i,j)$ is defined as another index to avoid the erroneous judgment when $u(i,j)$ is unable to separate the impulse noises. Suppose that the size of a sliding window is 3×3, and two pixel values; $s_1(i,j)$and $s_2(i,j)$ are defined as the closest value to the $x(i,j)$. Hence, the absolute difference between x(i, j) and these two pixels are obtained as a, b, that is a = $|x(i,j)- s_1(i,j)|$, b = $|x(i,j)- s_2(i,j)|$. Then, $v(i,j)$ denote the average of a and b, that is $v(i,j)$ = (a + b)/ 2. These indicators $u(i,j)$ and $v(i,j)$ are represented as large when impulsive noise is assumed existed, and vice versa.

Considering these both indicators, the isolation of impulsive noises is taken into consideration to separate the impulsive noises from the fine components of signals. In a filtering example, suppose that an image contains very fine components such as line components, *x*(*i,j*) is located on the line which is just one pixel width with no impulsive noise. Assume that the output of sliding window in median filter is taken as *m*(*i,j*), the value *u*(*i,j*) (or | *x*(*i,j*) - *m*(*i,j*)|) is large since *m*(*i,j*) is must not close to *x*(*i,j*), because an impulsive noise is assumed to be located at the pixel (*i,j*) to the background of this line, according to the indicator *u*(*i,j*). However, applied by the indicator *v*(*i,j*), since two selected input signals must be located on the line, then *v*(*i,j*) is derived as small, the inference system can tell that an impulsive noise is not located at the pixel (*i,j*).

If the combination of these rules is take as consideration, $\mu[\mathbf{w}(i,j)]$ can be expressed as a two-dimensional function of *u*(*i,j*) and *v*(*i,j*). In most fuzzy systems, each membership function is obtained by fuzzy reasoning using the membership function of the smallness and the largeness of *u*(*i,j*) and *v*(*i,j*) and its relationship to $\mu[\mathbf{w}(i,j)]$. Due to the difficulty at precisely the membership functions for *u*(*i,j*) and *v*(*i,j*), a method to design this membership function $\mu[\mathbf{w}(i,j)]$ is proposed by approximating this function by a step-like function, and setting the height of each step so that the mean square error of the filter output can be the minimum for training signal data.

Accordingly, suppose that the *u*(*i,j*) is includes in the *k*th piecewise region, *v*(*i,j*) is includes in the *m*th piecewise region (denote as in the (*k,m*) region latter), and the membership function derived from the pixel (*i,j*) is obtained as $\mu_{k,m}$. Hence, the output of the filter is depicted as

$$y(i,j) = m(i,j) + \mu_{k,m}(x(i,j) - m(i,j)) . \tag{2}$$

Consequently, the value of $\mu_{k,m}$ is obtained iteratively by a gradient method (the LMS algorithm) as follows,

$$\mu_{k,m}(n+1) = \mu_{k,m}(n) + \alpha(x(i,j) - m(i,j))e(i,j) , \tag{3}$$

where α is denoted as a convergence factor which is set based on the speed and the stability required in the convergence, *e*(*i,j*) is represented as an output error, and n is represented as the time point to process each pixel during a iteration. When the pixel (*i,j*) is processed, time point n is represented as the correspond time. Furthermore, when the next pixel locate in the (*i,j*) region again, the time point is processed as *n*+1. After $\mu_{k,m}$ converges, the value $\mu_{k,m}$ is obtained as the optimum solution to derive the minimum square error of the output.

3 The Fuzzy Neighboring System

3.1 Rotation Method

According to the previous statement, for the purpose to save the memory cost in fuzzy median filter, the nonlinear membership functions also can be used to replace step-like membership function, such as a Gaussian type or a sigmoid function, the form of the membership function is constrained to be close to these functions. Hence, the filter will be expressed as a nonlinear form of the parameters, the performance of the training result would be restricted and getting worse and the convergence speed can become very slow.

Thus, applying a step-like membership function, the constraint of the form of the membership function is less, and since the filter is expressed as a linear form of the parameter $\mu_{k,m}$ in each piecewise region, the performance of the training is as good as usual linear filter. However, the width of the steps in approximating the membership function as a step-like function is hard to design, since each axis is equally divided into piecewise regions with this width, and the size of divided area will influence the filter performance with contains of different steps for *u*(*i,j*) and *v*(*i,j*).

In this study, the author proposes a novel digital filter based on fuzzy median filter, neighboring fuzzy median filter, to improve the performance of the image restoration. In our study, the two-dimensional (2D) fuzzy model is used to represent the relations between the parameters *u*(*i*, *j*), *v*(*i*, *j*), and inference output $\mu[\mathbf{w}(i,j)]$. Inspecting the fuzzy model, the values in neighboring piecewise regions are similar to the value in (*k*, *m*) region. That is, there is compatibility among the neighboring piecewise regions. Therefore, using the structural features and object rotating method, our proposed filter can measure the neighboring outputs around the derived result to improve the filter performance. Under the restriction of the width of the steps in approximating the membership function, rotating the visual appearance of 2D fuzzy model is the extend method to obtain more possible solutions in local neighborhood, which refers to a number of equidistant pixels. Hence, more reliable outputs are available as an indicator to judge whether an impulse noises exist on those input pixels located in the same region.

Accordingly, a rotation method is applied as a visualization solution to derive the neighboring data. Given a rotation axis $\vec{n}=(n_x,n_y,n_z)$ in R^3 emanating from (0,0,0) and the rotation angle $\phi\in[-\pi,\pi]$, a corresponding rotation matrix *T* will be:

$$T=\begin{bmatrix} 1-2(n_y^2+n_z^2)\sin^2\frac{1}{2}\phi & -n_z\sin\phi+2n_xn_y\sin^2\frac{1}{2}\phi & n_y\sin\phi+2n_zn_x\sin^2\frac{1}{2}\phi \\ n_z\sin\phi+2n_xn_y\sin^2\frac{1}{2}\phi & 1-2(n_z^2+n_x^2)\sin^2\frac{1}{2}\phi & -n_x\sin\phi+2n_yn_z\sin^2\frac{1}{2}\phi \\ -n_y\sin\phi+2n_zn_x\sin^2\frac{1}{2}\phi & n_x\sin\phi+2n_yn_z\sin^2\frac{1}{2}\phi & 1-2(n_x^2+n_y^2)\sin^2\frac{1}{2}\phi \end{bmatrix}$$

Note that $\phi>0$ means the rotation angle is counterclockwise in relation to a viewpoint, which is along rotation axis looking toward the origin, and vice versa. Any point $A=(a_1,a_2,a_3)$ in R^3 is rotated according to a rotation axis $\vec{n}=(n_x,n_y,n_z)$ with angle ϕ will be $A'=(a_1',a_2',a_3')$, that is,

$$A' = T \cdot A \,. \tag{4}$$

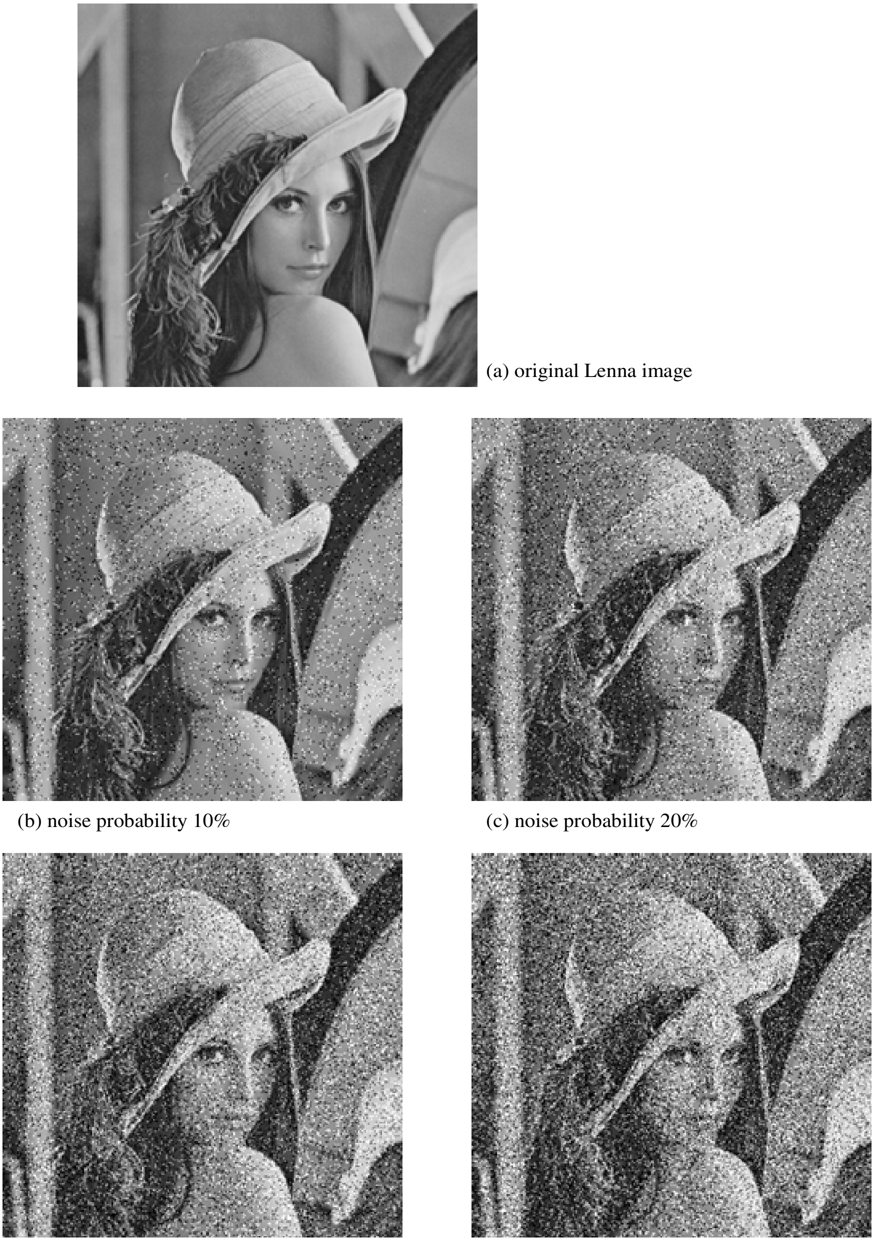
The rotation for the system outputs is the key of offer different $\mu[\mathbf{w}(i,j)]$ s for sliding windows with the same $u(i, j)$, $v(i, j)$ at region (k, m). It's included fuzzy median filter, the rotation method and the decision processing. The filter system works as follows, a sliding window is read from the polluted image as input. The usual fuzzy median filter puts out the estimated value, denoted as $\mu[\mathbf{w}(i,j)]$. Meanwhile, under the rotation mechanism, fuzzy median filter transfer the located region as a parameter, and the rotation method will rotate the region to derive the possible states, which is defined as $\mu[\mathbf{w}(i,j)]|_s$. After that, since the outputs of the system are finite elements of $\mu[\mathbf{w}(i,j)]|_s$ by rotation method and $\mu[\mathbf{w}(i,j)]$ by fuzzy median filter, a learning vector quantization (LVQ) network is designed to recognize these values as the best solution. Hence, the additional inputs of the LVQ network are the parameters $u(i,j)$, $v(i,j)$ and the standard deviation of sliding window, $t(i,j)$. Then, the output of the LVQ network is a best solution, which is defined as $\mu[x(i,j)]|_{\mathrm{R}}$.

3.2 LVQ System

The LVQ network is used to model the behavior of rotation in the system. The inputs of the LVQ network, which are the parameters $u(i, j)$, $v(i, j)$ and the standard deviation $t(i, j)$, are collected from a sliding window of an image. The output of the LVQ network is a parameter selected form the subset $\mu|x(i,j)||_{\mathrm{R}}$ which is collected under the rotation of the set $\mu|x(i,j)|$. The LVQ network is based on the fuzzy relation. The key of the LVQ training is how to confirm the correctness of classification. The principle of judging the correctness of classification in training process is to determine the best $\mu|x(i,j)||_{\mathrm{R}}$ under the rotation for each region (k, m) before LVQ training. The region (k, m) will 'meet' finite elements of $\mu|x(i,j)|$ under rotation. A LVQ network is designed to recognize what the value of $\mu|x(i,j)||_{\mathrm{R}}$ should be classified from the input vector with $u(i, j)$, $v(i, j)$ and $t(i, j)$ at region (k, m) by rotation. It is unreasonable for a single LVQ network, or any single neural network to tell what the patterns $\mu|x(i,j)|$ might occur after the rotation at different regions (k, m) because it is too complex. Thus, if the membership function $u(i, j)$ divided into r piecewise region and $v(i, j)$ into s piecewise region, then it is defined there are t ($=r\times s$) LVQ networks to model the behavior of rotation for different cases (k, m).

4 Experiment Result

Since the mechanism of rotation is based on the fuzzy system, the performance of the FMF on image restoration is presented first and then the system with the rotation mechanism is shown. The image shown in Fig. 1 (a) is an original Lenna image that is used to construct the fuzzy system for application. The picture format of Fig. 1 is GIF, the size of image is 256 by 256 and the value of gray level is between 0~255. Four images with distinct impulse noises probabilities added are shown in Figs.1 (b)~(e) which are noisy images going to been dealt with.



(a) original Lenna image

(b) noise probability 10%

(c) noise probability 20%

(d) noise probability 30%

(e) noise probability 40%

Fig. 1. The testing Lenna image with distinct noise probability

Table 1 shows the MSE of the output of a conventional median filter. It is obviously that the more the noise exists, the performance of median filter is worse. Tables 2 displays the MSE of FMF. The membership function $\mu[x(i, j)]$ via $u(i, j)$ and $v(i, j)$ of FMF is trained by the original images of Lenna shown for different cases. The initial value of $\mu[x(i, j)]$ is all set to 0.2, $\mu_{k,m}$ remain 0.2 when no sliding windows are applied in the (k, m) regions. The index $n \times n$ indicates that each step-like function of $u(i, j)$ and $v(i, j)$ is divided into n piecewise regions. A larger n indicates a smoother step-like function. The performance of fuzzy-median filter is better than the conventional median filter. The fact is shown by comparing the MSE of Table 1 and Tables 2. The smoother the step-like function is, the better performance of fuzzy-median filter will be.

Table 1. The MSE of the output of conventional median filter

	10%	20%	30%	40%
Lenna	67.17	91.24	142.66	234.90

Table 2. The MSE of the FMF for Lenna image

Lenna	10%	20%	30%	40%
5 * 5	37.23	67.68	119.31	213.45
10 * 10	29.54	59.32	110.73	203.83
25 * 25	26.14	53.56	105.32	198.92
50 * 50	24.22	50.75	101.35	194.96

5 Conclusion

A formal and systematical method applying the human language to the computer program is constructed by knowledge representation. Knowledge can be represented by fuzzy logic and a rule-based system. A fuzzy system constructed by fuzzy rule base and fuzzy inference engine based on fuzzy logic translates the knowledge or information about the world into computer program. The humanization of a fuzzy system by the mechanism of “cudgel one’s brains” is proposed in this thesis by rotating the knowledge structure of the fuzzy system. The intelligent mechanism is based on rotation. The human behavior “cudgel one’s brains” is searching some similar concept or knowledge for better solutions. The intelligent mechanism is based on this idea and realized by rotating the knowledge described by a fuzzy system. A FMF is proposed to improve the conventional median filter, where the knowledge about the relation of impulse noises, line components and how to process a current mask has been trained by LMS. By using step-like membership function, masks at the same region take signal value on solving problem. The intelligent mechanism offers choices for theses masks by rotating the knowledge. The point of view is proved by extensive experiments. The results show that the performance of FMF with the intelligent mechanism is better than original FMF on noisy image restoration.

References

1. Chen, T., Wu, H.R.: Space Variant Median Filters for the Restoration of Impulse Noise Corrupted Images. IEEE Trans. Circuits Syst. –II: Analog and Digital Signal Processing 48, 784–789 (2001)
2. Astola, J., Kuosmanen, P.: Fundamentals of Nonlinear Digital Filtering. CRC, Boca Raton (1997)
3. Jain, A.K.: Fundamentals of Digital Image Processing. Prentice-Hall, Englewood Cliffs (1989)
4. Chen, T., Wu, H.: Adaptive Impulse Detection Using Center-Weighted Median Filters. Signal Processing Lett. 8(1), 1–3 (2001)
5. Sun, T., Neuvo, Y.: Detail-preserving Median Based Filters in Image Processing. Pattern Recognit. Lett. 15(4), 341–347 (1994)
6. Abreu, E., Lighstone, M., Mitra, S.K., Arakawa, K.: A New Efficient Approach for the Removal of Impulse Noise from Highly Corrupted Images. IEEE Trans. Image Processing 5, 1012–1025 (1996)
7. Arakawa, K.: Median Filter Based on Fuzzy Rules and Its Application to Image Restoration. Fuzzy Sets and Systems 77, 3–13 (1996)
8. Yan, Z.: Adaptive Fuzzy Median Filter for Images Corrupted by Impulsive Noise. In: Congress on Image and Signal Processing (2008)
9. Sukomal, M., Sanjeev, D.: Fuzzy Based Median Filter for Gray-scale Images. International Journal of Engineering Science & Advanced Technology 2, 975–980 (2012)

An Elastic Net Clustering Algorithm for Non-linearly Separable Data

Chun-Wei Tsai[1], Chien-Hung Tung[2], and Ming-Chao Chiang[2]

[1] Information Technology, Chia Nan University of Pharmacy & Science
Tainan, Taiwan
cwtsai0807@gmail.com

[2] Computer Science and Engineering, National Sun Yat-sen University
Kaohsiung, Taiwan
tim30235@yahoo.com.tw, mcchiang@cse.nsysu.edu.tw

Abstract. This paper presents an effective elastic net-based clustering algorithm for complex and non-linearly separable data. The basic idea of the proposed algorithm is simple and can be summarized into two steps: (1) assign patterns to groups based on the attraction and tension between the elastic nodes in a ring and neighbors of the patterns and (2) merge the groups based on the distance between the elastic nodes. To evaluate the performance of the proposed method, we compare it with several state-of-the-art clustering methods in solving the data clustering problem. The simulation results show that the proposed algorithm can provide much better results than the other clustering algorithms compared in terms of the accuracy rate. The results also show that the proposed algorithm works well for complex datasets, especially non-linearly separable data.

Keywords: Data clustering, *k*-means, elastic net.

1 Introduction

Clustering [1,2] is an important data analysis technology, for it can be used to mine information hidden in a dataset and it has been widely used in several different problem domains. Also, many successful cases in text, image, video, and voice mining, bioinformatics, risk management of credit card, or sunspot prediction [3,4,5,6] show the potential of clustering being a critical data analysis technology today and in the future.

The partitional clustering problem [7,8] is defined as classifying (or partitioning) a set of input patterns $X = \{x_1, x_2, \ldots, x_n\}$ in the r-dimensional space into k distinct clusters $\Pi = \{\pi_1, \pi_2, \ldots, \pi_k\}$ represented by a set of means or centroids $C = \{c_1, c_2, \ldots, c_k\}$ in the same space where $\pi_i = \{x \in X \mid d(x, c_i) \leq d(x, c_j), \forall i \neq j\}$, $c_i = \frac{1}{|\pi_i|} \sum_{\forall x \in \pi_i} x$, and k is usually predefined by the user or automatically determined by the clustering process. The similarity metric $d(\cdot, \cdot)$ is used to measure the quality of the clustering results. It somehow depends on the problem in question.

The development of modern clustering algorithms to solve clustering problems from different problem domains dates back to the 1950s or even earlier.

A. Selamat et al. (Eds.): ACIIDS 2013, Part I, LNAI 7802, pp. 108–116, 2013.

Among them are k-means [9], genetic algorithm (GA) [10], and particle swarm optimization (PSO) [11]. Although these clustering technologies are very useful for clustering problems; they are, however, usually only designed for particular problems instead of for all problems. According to the observation of [1,2], most clustering algorithms suffer from the follows dilemmas: (1) they perform poorly for large datasets, (2) they are sensitive to the initial seeds, (3) they cannot automatically determine the number of clusters, (4) they usually converge to local minima, and (5) they perform poorly in clustering non-linearly separable data. For the last dilemmas, several researches have attempted to design an effective algorithm to deal the last dilemma, such as kernel-based clustering algorithm [12].

However, most kernel-based clustering algorithms suffer from the computationally expensive problem. This paper presents a simple but effective algorithm for clustering non-linearly separable data. The proposed algorithm leverages the strengths of the elastic net algorithm, by considering not only the relationships between patterns and centroids[1] but also the attraction and tension between the rubber band, the elastic nodes, and the input patterns. Thus, the proposed algorithm is able to recognize and differentiate the non-linearly separable data groups from a complex data distribution.

The remainder of the paper is organized as follows: Section 2 briefly discusses the elastic net algorithm and how to apply it to the clustering problem. Section 3 describes in detail the proposed algorithm. The data sets we evaluated and the parameter settings are given in Section 4. The experimental results are also discussed in this section. Section 5 concludes the work.

2 Related Work

To simplify our discussion of the elastic net algorithm, the following notations are used in this paper.

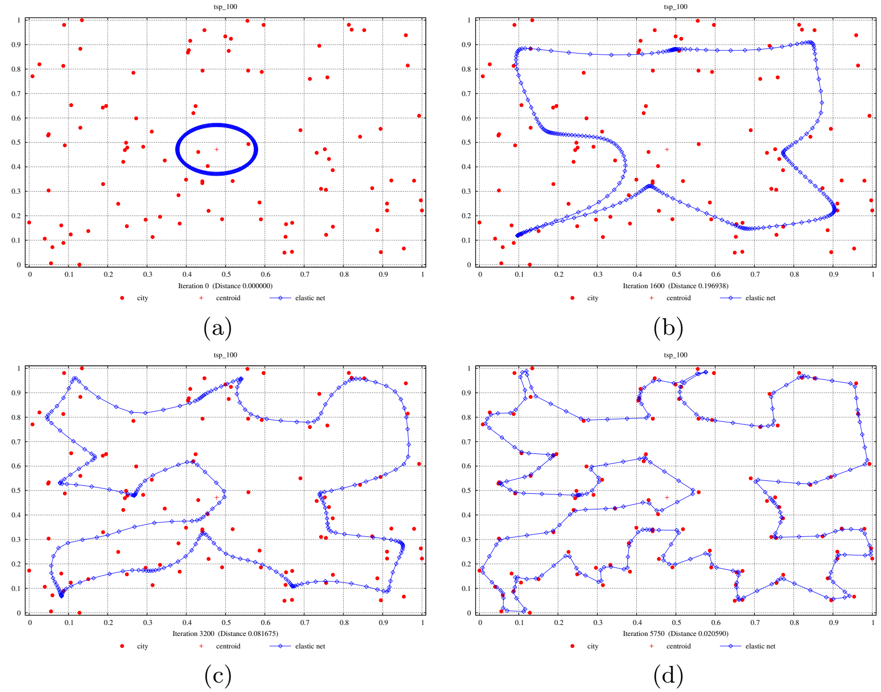
X	The set of data pattern, i.e., $X = \{x_1, x_2, \ldots, x_n\}$.
Y	The set of nodes on an elastic ring, i.e., $Y = \{y_1, y_2, \ldots, y_m\}$.
α	The coefficient of force between data patterns and nodes on the elastic ring.
β	The coefficient of force between nodes on the elastic ring and its neighbors.
K	A scale parameter which dominates the number of iterations.
w_{ij}	A normalization function to equalize the influence of all patterns.

2.1 Elastic Net Algorithm

The elastic net algorithm was first presented in [13,14] to create approximate solutions for the traveling salesman problem. The basic idea is to use a ring as a

[1] Most clustering algorithms that rely on the distance information between centroids and patterns are radius-based, thus not able to classify non-linearly separable data.

rubber band and sprinkle the so-called elastic nodes on it to extend the rim until it reaches all the nodes on the sample space. For the rubber band always forms a closed path, the lines will not be superimposed or crossed. Thus, the elastic net algorithm can be intuitively applying to the traveling salesman problem to avoid creating crossed routing paths of a tour.



(a)

(b)

(c)

(d)

Fig. 1. A simple example illustrating how the proposed algorithm works

The elastic net is as specified by Eq. (1). The idea is to move the points on the ring representing the force of attraction and tension of each elastic node (point) y_i at each iteration. In other words, $(x_i - y_j)$ represents the force between the input pattern (node or city) x_i and the elastic node y_j while $(y_{j+1} - 2y_j + y_{j-1})$ represents the force between the neighbors of an elastic node on the ring.

$$\Delta y_j = \alpha \Sigma_i w_{ij}(x_i - y_j) + \beta K(y_{j+1} - 2y_j + y_{j-1}), \tag{1}$$

where x_i denotes the position of i-th pattern (also called sample point) and y_j denotes the j-th elastic node on the rim. Also, Δy_j plays the role of changing the value of y_j at each iteration. w_{ij}, α, and β are used to weigh the attraction and tension of elastic nodes. The coefficient w_{ij}, defined as

$$w_{ij} = \phi(|x_i - y_j|, K)/\Sigma_k \phi(|x_i - y_k|, K), \tag{2}$$

where $\phi(d, K) = \exp(-d^2/2K^2) > 0$, represents the influence of the connection between x_i and y_j; α and β are predefined constants. K is a reduction of the length parameter.

Fig. 1 uses a 100-node traveling salesman problem [15] as an example. Fig. 1(a) shows that the EN will initially set a ring in the center of the input patterns. Fig. 1(b) and (c) show at iterations 1,600 and 3,200, the results are getting closer and closer to to a complete (legal) solution for the traveling salesman problem (TSP). Finally, at iteration 5750, the ring of EN fits to most of the cities, thus producing a non-cross solution for the TSP.

2.2 Elastic Net for Data Clustering

Because the EN method can provide a high search efficiency for complex combinatorial optimization problem, our observation shows that it is not only successfully used in solving the traveling salesman problem in the past, but is can also be used in solving complex clustering problem. The fuzzy elastic clustering (FECL) [16] is a typical approach which combines the fuzzy set with EN to provide better clustering results. However, because the FECL is not crisp, it needs to be defuzzified. Then, all the patterns may be classified into one cluster. To determine to which clusters patterns belong, the membership function membership(t) of FECL is defined as

$$\text{membership}(t) = \begin{cases} 1, & \text{if } d \leq r, \\ \exp\left(\frac{(d-r)^2}{r\sigma}\right), & \text{otherwise,} \end{cases} \tag{3}$$

where d is the distance between the cluster centers, r is the distance between p (the point on the ring that is closest to the test pattern t, as shown in Fig. 2), and σ is a parameter of the influence of r for the membership function. In other words, Fig. 2 gives a simple example to show how the FECL uses Eq. (3) to determine the membership values for assigning the patterns to the clusters they belong. The experimental results in [16] show that the accuracy rate of FECL is significantly affected by the parameters K and σ because they determine the convergence speed and the clustering strategy.

Another elastic algorithm for hierarchical clustering was presented in [17] for data in the n-dimensional space, called elastic neural net algorithm (ENNA), to solve the problem of the essence of EN being designed for the bi-dimensional space. More precisely, the ENNA consists of two stages: (1) each pair of nodes will be removed when they are closer than the other nodes and the ENNA will add a new node halfway between these two nodes and (2) associate all the nodes with the elastic band to further construct the final solution.

3 The Proposed Algorithm

3.1 Design and Implementation

The proposed algorithm (elastic net based clustering algorithm; ENCA) builds directly on the elastic net algorithm to solve the partitional clustering problem

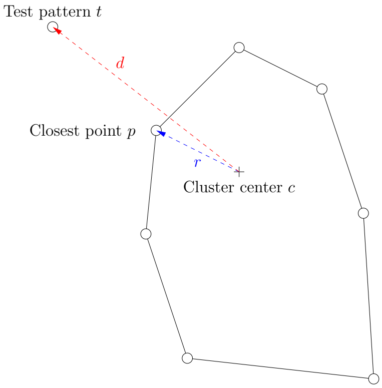


Fig. 2. Membership value as a function of distance

to avoid the problem that most clustering algorithms cannot deal with non-linearly separable data, such as curvilinear clusters. The basic idea is that the elastic nodes are employed to depict roughly the distribution of data; then, the proposed algorithm can use this information to classify the input patterns into approximate clusters. The main difference between the ENCA and the other elastic net-based clustering algorithms is in that the proposed algorithm uses two novel clustering operators to assign the patterns to the elastic nodes and the number of clusters will be merged from $k' \leq n/2$ down to k.

```
Initialize the parameters.
Generate an elastic ring and put it at the starting position.
Perform the simple elastic net algorithm.
Calculate the threshold $T_m$.
/* The first merge process */
Merge clusters the distance between them is less than $T_m$.
Let $k'$ denote the number of clusters left.
/* The second merge process */
While ($k' > k$) {
    Merge the pair of clusters with the smallest sum of the smallest distances between
        each elastic node in them.
    $k' \leftarrow k' - 1$.
}
Output the clustering results.
```

Fig. 3. The proposed algorithm ENCA

Fig. 3 gives an outline of the proposed algorithm, which works just as how most iterative-based clustering algorithms do. First, the number of elastic nodes m is set equal to $m = n/2$ where n is the number of input patterns. The radius of the rubber band is set equal to the average distance between all the input patterns to the centroid c.[2] Initially, the elastic ring is imposed on the least

[2] At this stage, all the patterns are assumed to belong to the same cluster; thus, there is one and only one centroid, which is referred to as c.

densely populated two dimensions of the input patterns; i.e., the two dimensions with the largest and second largest distances to the centroid c. The purpose is to avoid the possibility of the elastic ring not being able to be expanded by the proposed algorithm. Next, the elastic net process is performed to build up the distribution of the input patterns by using Eq. (1). Here, the value of K is updated once every 25 iterations by $K_t = 0.99K_{t-1}$; that is, $K = K_0$, $K = K_1 = 0.99K_0$, $K = K_2 = 0.99K_1$, and so on. Note that the decrease rate of K will affect the convergence speed.

In addition, to avoid the overflow problem due to the fact that the EN was originally designed only for 2-dimensional data instead of for data the dimension of which is larger than 2, the distance metric is defined as $d' = d/\sqrt{|r|}$, which is used to replace the Euclidean distance metric of the proposed algorithm on line 3 of Fig. 3. The proposed algorithm can then be safely used on data the dimension of which is higher than 2. Then, as line 3 of Fig. 3 shows, the elastic ring will take the shape as a cluster of all the input patterns; i.e., all the patterns are assigned to the closest elastic node. However, in this step, some of the elastic nodes will not be assigned any pattern. The proposed algorithm will then remove these elastic nodes because they are not assigned any pattern. Now, each elastic node represents a cluster, and the number of clusters is smaller than $n/2$ because some elastic nodes are not assigned any pattern.

After that, the threshold T_m as shown in line 4 of Fig. 3 is computed as follows: First, the distances between all elastic nodes are computed and sorted. Let us denote them by $D = \{d_1, d_2, \ldots, d_h\}$. Next, the average μ and standard deviation σ of all the distances in D are computed, and everything in D that is larger than $\mu + 2\sigma$ is removed. Let us call the set thus formed D'. Then, the average μ' of all the distances in D' is computed, and everything in D' that is smaller than μ' is removed. Finally, the average of all the distances left over in D' is computed, which is the threshold T_m we are seeking.

As shown in line 6, the first merge process is performed on all clusters. That is, the proposed algorithm will merge each pair of clusters the distance between them is less than T_m. Finally, as shown in the loop beginning at line 9, all the remaining clusters will be merged one by one until the number of clusters is equal to k. Here is how it works. First, the sum of the smallest distances between each elastic node in one cluster and each elastic node in another is computed. Then, the pair of clusters with the smallest sum is merged.

3.2 An Example

4 Empirical Analysis

4.1 The Simulation Environment and Parameter Settings

To evaluate the performance of the ENCA for the clustering problem, the proposed algorithm is compared with k-means [9], genetic k-means algorithm (GKA) [18], and K-means with genetic algorithm (KGA) [19].

The empirical analysis was conducted on an ASUS i7 machine with 2.67 GHz i7-920 CPU and 4GB of memory using Fedora 12 running Linux 2.6.31. Moreover, all the programs are written in C++ and compiled using g++ (GNU C++ compiler). Table 1 gives the benchmarks used in this paper. All the simulations are carried out for one run, with the parameters α set equal to 0.2; β set equal to 2.0, and K set equal to 0.24.

Table 1. Dataset

Dataset	# patterns	# dimensions	# clusters
Synthetic	150	2	3
WBC	699	9	2
Iris	150	4	3
Wine	178	13	3
Ecoli	336	7	8
Haberman	306	3	2
Glass	214	9	6

4.2 The Simulation Results

The simulation results show that the other clustering algorithms we compared in this paper outperforms the ENCA in terms of SSE (sum of squared error). However, in terms of the accuracy rate (AR), the proposed algorithm beats the other clustering algorithms for most of the datasets. According to our observation, this is because most clustering algorithms use the centroids to represent the clustering results and the radius-based method to differentiate and determine the clusters to which the patterns belong. It makes most of the cluster algorithms unable to handle non-linearly separable data because the data are not circularly distributed. In other words, if the dataset is composed of curvilinear clusters, these clustering algorithms may erroneously assign the patterns into wrong clusters. To overcome this problem, the proposed algorithm is composed of two steps. The first step is aimed at finding out the distribution of data while the second step is aimed at combining the clusters found in the first step,

Table 2. The results of ENCA, *k*-means [9], GKA [10], and KGA [19]

Dataset	ENCA		*k*-means		GKA		KGA	
	AR	SSE	AR	SSE	AR	SSE	AR	SSE
Synthetic	**100.00%**	16.752	76.00%	11.991	75.55%	11.984	73.57%	12.726
WBC	**96.42%**	244.811	95.68%	237.996	95.78%	240.124	95.63%	238.591
Iris	**96.67%**	7.993	82.26%	7.808	87.30%	7.432	84.51%	7.6212
Wine	92.69%	50.321	**94.53%**	48.984	94.32%	48.962	92.10%	50.607
Ecoli	**78.87%**	20.074	54.20%	19.371	61.52%	20.048	54.89%	18.816
Haberman	**59.48%**	28.481	51.29%	25.322	52.61%	25.280	51.52%	25.599
Glass	**49.53%**	22.943	43.42%	20.445	43.36%	18.998	42.18%	21.573

by using two merge operators one of which uses the global distance information (i.e., T_m as the threshold to determine which two clusters should be merged) to merge clusters while the other of which uses the local distance information (i.e., distance between two groups). In other words, the proposed algorithm takes into consideration not only the distances between the patterns and the centroids to which they belong but also the geographical distribution of the input dataset. That is why the proposed algorithm can provide better results than the other traditional clustering algorithms.

5 Conclusion

This paper presented a novel elastic net algorithm to solve the problem of clustering non-linearly separable data. With a simple modification to the EN, the proposed algorithm can be used to cluster data the dimension of which is higher than 2. Two novel merge operators are proposed for the proposed algorithm to prevent the noise patterns from affecting the clustering results of the elastic net-based clustering algorithm. Also, Not only is the convergence speed of the proposed algorithm faster than the EN because the number of elastic nodes is reduced, it also provides a better result than the other heuristic clustering algorithms. In the future, our goal is to focus on finding an even more efficient operator to enhance the quality of the final result.

Acknowledgment. This work was supported in part by the National Science Council of Taiwan, ROC, under Contracts NSC101-2221-E-041-012 and NSC99-2221-E-110-052.

References

1. Jain, A., Murty, M., Flynn, P.: Data clustering: A review. ACM Computing Surveys 31(3), 264–323 (1999)
2. Xu, R., Wunsch-II, D.C.: Survey of clustering algorithms. IEEE Transaction on Neural Netoworks 16(3), 645–678 (2005)
3. Leuski, A.: Evaluating document clustering for interactive information retrieval. In: Proceedings of the International Conference on Information and Knowledge Management, pp. 33–40 (2001)
4. Xu, W., Liu, X., Gong, Y.: Document clustering based on non-negative matrix factorization. In: Proceedings of the International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 267–273 (2003)
5. Getz, G., Gal, H., Kela, I., Notterman, D.A., Domany, E.: Coupled two-way clustering analysis of breast cancer and colon cancer gene expression data. Bioinformatics 19, 12079–12084 (2003)
6. Xu, R., Wunsch-II, D.C.: Clustering. Wiley, John & Sons, Inc. (2008)
7. Theodoridis, S., Koutroumbas, K.: Chapter 16: Cluster Validity. In: Pattern Recognition, 4th edn., pp. 863–913. Academic Press, Boston (2009)
8. Xiang, S., Nie, F., Zhang, C.: Learning a mahalanobis distance metric for data clustering and classification. Pattern Recognition 41(12), 3600–3612 (2008)

9. McQueen, J.B.: Some methods of classification and analysis of multivariate observations. In: Proceedings of the Berkeley Symposium on Mathematical Statistics and Probability, pp. 281–297 (1967)
10. Maulik, U., Bandyopadhyay, S.: Genetic algorithm-based clustering technique. Pattern Recognition 33(9), 1455–1465 (2000)
11. Omran, M.G., Salman, A.A., Engelbrecht, A.P.: Image classification using particle swarm optimization. In: Proceedings of the Asia-Pacific Conference on Simulated Evolution and Learning, pp. 370–374 (2002)
12. Dhillon, I.S., Guan, Y., Kulis, B.: Kernel k-means: spectral clustering and normalized cuts. In: Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 551–556 (2004)
13. Durbin, R., Willshaw, D.: An analogue approach to the travelling salesman problem using an elastic net method. Nature 326, 689–691 (1987)
14. Durbin, R., Szeliski, R., Yuille, A.: An analysis of the elastic net approach to the traveling salesman problem. Neural Computation 1, 348–358 (1989)
15. TSPLIB (2012), http://www.iwr.uni-heidelberg.de/groups/comopt/software/TSPLIB95/tsp/
16. Srikanth, R., Petry, F.E., Koutsougeras, C.: Fuzzy elastic clustering. In: Proceedings of the IEEE International Conference on Fuzzy Systems, vol. 2, pp. 1179–1182 (1993)
17. Salvini, R.L., de Carvalho, L.A.V.: Elastic neural net algorithm for cluster analysis. In: Proceedings of the Brazilian Symposium on Neural Networks, pp. 191–195 (2000)
18. Krishna, K., Murty, M.N.: Genetic k-means algorithm. IEEE Transactions on System, Man and Cybernetics—Part B:Cybernetics 29(3), 433–439 (1999)
19. Bandyopadhyay, S., Maulik, U.: An evolutionary technique based on k-means algorithm for optimal clustering in R^N. Information Sciences 146(1-4), 221–237 (2002)

Anticipatory Emergency Elevator Evacuation Systems

Kai Shi[1,2], Yuichi Goto[1], Zhiliang Zhu[2], and Jingde Cheng[1]

[1] Department of Information and Computer Sciences, Saitama University
Saitama, 338-8570, Japan
{shikai,gotoh,cheng}@aise.ics.saitama-u.ac.jp
[2] Software College, Northeastern University
Shenyang, 110819, China
{shik,zhuzl}@swc.neu.edu.cn

Abstract. This paper proposes a new type of *emergency elevator evacuation systems*, named *anticipatory emergency elevator evacuation systems*, which can detect and predict hazards in the emergency evacuation, and then dispatch the elevator cars anticipatorily, aiming to avoid hazards beforehand, thus to rescue more occupants and to shorten the evacuation time. This paper presents the proposed system's design, its prototype and evaluation. The novelty of the presented work is that it shows a new direction for developing emergency elevator evacuation systems.

Keywords: anticipatory computing, context-aware computing, anticipatory reasoning-reacting system.

1 Introduction

Using elevators for emergency evacuation has become reasonable and necessary for modern high-rise buildings in recent years. First, it is impossible to evacuate an ultra high-rise building in a tolerable time by only using stairs [1]. Using elevators can decrease evacuation time and reduce congestion on stairs which equates to less potential for injuries [2]. Second, the aged and people with disabilities or injuries can hardly use stairs to evacuate [1]. Third, fire is not the only reason for evacuation. Sometimes it is important to leave the building before a situation gets worse, such as a bomb threat or other acts of terrorism [2]. Lastly, the current elevator systems can provide safe and reliable operation both for fire service access and for occupant egress during fires [2, 3].

An *emergency elevator evacuation system* (EEES) includes the elevator equipment, hoistway, machine room, and other equipment and controls needed for safe operation of the elevator during the evacuation process [4]. Because the purpose of emergency evacuation is to move people away from the threat or actual occurrence of a hazard immediately and rapidly, an ideal EEES should have the following features: safe, context aware, anticipatory, and instructional/informative. However, the current researches of EEESs mainly focus on the physical safety of the elevator systems; few researches involve context aware of EEESs; let alone anticipation.

As an attempt to realize an advanced EEES with safe, context aware, anticipatory, and instructional/informative features, this paper proposes a new type of EEESs, named *anticipatory emergency elevator evacuation system* (AEEES), which can detect and

A. Selamat et al. (Eds.): ACIIDS 2013, Part I, LNAI 7802, pp. 117–126, 2013.

predict hazards in the emergency evacuation, and then dispatch the elevator cars anticipatorily aiming to avoid hazards beforehand, thus to rescue more occupants and to shorten the evacuation time. In order to implement the AEEES, we adopted anticipatory reasoning-reacting system (ARRS) [5] as the control system of the AEEES. ARRS's key function is its ability of *anticipatory reasoning*, which can draw new, previously unknown and/or unrecognized conclusions about some future event or events whose occurrence and truth are uncertain at the point of time when the reasoning is being performed [5]. To evaluate the AEEES, this paper presents experiments for comparison the basic EEES which dispatch elevator cars using down-peak [6], the reactive EEES with context awareness but without anticipation, and our proposed AEEES. However, this paper do not discuss the physical safety of the EEES, because there are a lot safety elevator have been practically used [1].

2 Ideal Emergency Elevator Evacuation Systems

What should an ideal *emergency elevator evacuation system* (EEES) be? Because the purpose of emergency evacuation is to move people away from the threat or actual occurrence of a hazard immediately and rapidly, an ideal EEES should have the following features.

Safe: The EEES must ensure the safety of the passengers who take the elevators during the emergency evacuation.

Context aware: The EEES can autonomically both sense and react against the particular case of emergency. "Sense" means the EEES can detect the emergency events (e.g., fire, gas leaks), perceives the current situation (e.g., which rooms are catching fire in a fire emergency, an elevator car is full loaded), and recognizes special people (e.g., people with disabilities or injuries). "React" means the EEES decides which evacuation type is taken according to the emergency situation (including total evacuation, staged evacuation, and fractional evacuation [2]), and reacts to some particular events (e.g., not stopping the elevator car in fire, dispatching a fully loaded elevator directly to the evacuation floor).

Anticipatory: In order to indeed make "people away from the threat", the EEES should be anticipatory, i.e., the EEES should predict the future situations (e.g., who will be in danger soon, and there will be a congestion in some stairs), then dispatch the elevator cars beforehand, aiming to avoid disaster beforehand and shorten the evacuation time. A serious emergency usually gets worse rapidly with deadly damage. For example, in an uncontrollable conflagration, because the fire spread with great rapidity, it is too late to egress when fire and poisonous smoke draw near. Thus the EEES should have the ability to predict which area of building will be dangerous, then rescue the people in that area. Besides, if the EEES can predict where a congestion will happen, then transport some people in that area, thus a trample may be avoided.

Instructional/informative: The EEES can instruct occupants to use the elevators to evacuate, and inform occupants the current situation of emergency by public address or other approaches. Besides, the EEES must instruct occupants which elevators can be used, who should use the elevator first, as well as inform occupants the emergency situation, and how long they have to wait the next shift.

3 Current Emergency Elevator Evacuation Systems

The current researches of emergency elevator evacuation systems (EEES) mainly focus on the physical safety of the elevator systems, and there is still a gap between current EEESs and ideal EEESs. Klote et al. studied the feasibility of elevator evacuation of FAA air traffic control towers and developed the concept of EEES [4]. The researches of EEESs can be divided two categories: physical safety and elevator cars dispatch for emergency.

Physical safety is the basic challenge of EEESs, i.e., to protect elevators from heat, flame, smoke, water, overheating of elevator machine room equipment, loss of electric power, as well as to assure the safety of people traveling in the elevators. Several researches were carried out on this topic [3, 4, 7]. As a result, several kinds of elevators for evacuation were developed, which can provide the above safety features, such as the enhanced elevators and protected elevators [2]. By now, some skyscrapers have equipped these elevators for evacuation [1].

In addition to safety, another challenging problem for EEES is how to control elevators autonomically, effectively and safely according to the current evacuation type and extraordinary event. Until now, there are mainly following elevator control strategies for evacuation.

Neglect: Some EEESs do not provide additional control strategies for evacuation. These systems just use general elevator algorithm or general group control algorithm for evacuation.

Manual control and semiautomatic control with human oversight [8, 9]: Large buildings often have a command center for directing an emergency evacuation. Operators in command center can get information from alarm systems, floor monitors, television or security cameras, then dispatch elevators to where they are needed manually. Operators can also use some automated control system to set priorities and determine which floors should evacuate using the elevators.

Down-peak and its variation: In down-peak mode, elevator cars depart from the lobby to the highest floor served, then run down the floors in response to hall calls placed by passengers wishing to leave the building. There are also some improved group elevator down-peak scheduling (including zoning of elevator groups) for emergency evacuation [10]. For evacuation, down-peak mode is proved more better than other special operating modes such as up-peak [6]. The benefit of down-peak is easy to implement. However, only using down-peak cannot consider different emergency evacuation types, such as fractional emergency evacuation [2], and current perils and situation of the scene.

Dispatching elevators based on the current situation: Some patents claimed their systems can measure the number of people remaining inside a building, detect an emergency condition in the building, then dispatch the elevators according to these current situation [11, 12, 13]. However, these patents do not show how to fulfill their claim in detail. Moreover, these systems can only dispatch elevators based on certain kinds of information, but cannot deal with complex situation dynamically.

Theoretical evacuation strategies: There are some theoretical elevator evacuation strategies were proposed, that can be used to determine which floor should be evacuate firstly [14, 15, 16]. However, they are only theoretical work and do not show how to use these strategies in a practical EEES.

In summary, although the current EEESs have made big progress in their physical safety, the current EEESs are still in an early phase for autonomic evacuation elevator control, because the current EEESs 1) cannot automatically deal with different evacuation types, different extraordinary events, and cannot dispatch the elevator cars according to the evacuation type, the extraordinary event, and the current situation, and 2) does not consider anticipation.

4 Anticipatory Emergency Elevator Evacuation Systems

This paper proposes a new type of EEESs, named *anticipatory emergency elevator evacuation system* (AEEES). An AEEES can automatically detect and predict hazards in the emergency evacuation, and then dispatch the elevator cars anticipatorily, aiming to avoid disasters beforehand, thus to rescue more occupants and to shorten the evacuation time. Because the physical safety of AEEES is ensured by the enhanced/protected elevators, the heart of AEEES is an information system to control elevator cars and to instruct occupants. Thus, this paper only focuses on the information system, but not the whole of AEEES as well as the physical safety.

In order to fulfill the features of context aware, anticipatory, and instructional/informative, we analyzed the system requirements of the AEEES as follows.

R1: The AEEES should perceive the current situation of the building by utilizing the sensory data from different kinds of sensors and monitor systems.

R2: The AEEES should deal with different type of emergency, such as fire, gas leaks, bomb threats, as well as several types of emergency happens in the same time. Besides, the system should deal with new type of emergency with trivial modification.

R3: The AEEES should decide the emergency type according to the current emergency, including total evacuation, staged evacuation, and fractional evacuation [2].

R4: The AEEES should predict the future hazards automatically and autonomically .

R5: The AEEES should automatically generate anticipatory actions (against the future hazards), reactive actions (against the ongoing hazards), and routine actions (for total occupation, e.g. down peak). Besides, the AEEES should judge which actions should be taken first automatically.

R6: The AEEES should inform the corresponding occupants the ongoing hazards and the predictions of hazards, and which elevators can be used and when next shift comes, as well as instruct the occupants what to do next. The content of the messages are chosen automatically. The AEEES gives messages by public address system or other communication systems.

R7: The AEEES should know evacuation for certain type of emergency should stop automatically, when that emergency was eliminated (e.g. the fire dies out).

R8: The AEEES should be portable for different buildings with trivial modification, such as different building structure, different elevator systems, different sensors, and different layout of elevators and sensors.

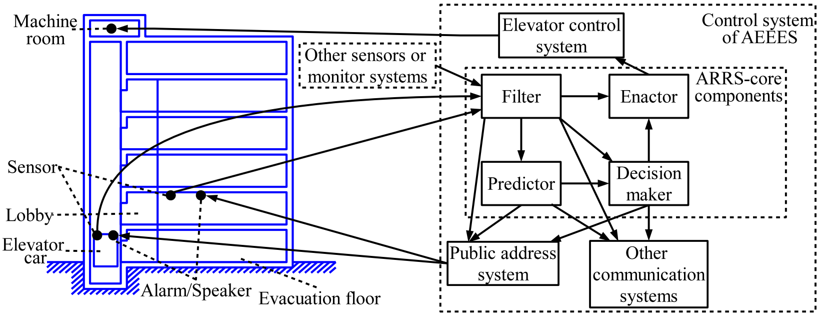


Fig. 1. An architecture of ARRS-based AEEES

5 An Implementation Based on Anticipatory Reasoning-Reacting System

In order to realize prediction and decision making of the AEEES, we adopted anticipatory reasoning-reacting system (ARRS) [5] as the control system of the AEEES. An ARRS uses logical reasoning to predict and make decisions [5]. The basic idea of logical reasoning method is to explicitly separates the underlying logic system, reasoning/computing mechanism, and empirical knowledge in any prediction/decision making, such that both underlying logic system and empirical knowledge can be revised/replaced/customized in various predictions/decision making processes performed by an area-independent, task-independent, general-purpose reasoning mechanism. Therefore, the generality of logical reasoning method is fit to realize requirements **R2** and **R8**. An ARRS consists of traditional reactive systems and core components, which include predictor, decision maker, filter, and enactor. Such an architecture is fit to embrace other reactive system such as the elevator system. Due to space limitations, this paper only present briefly about prediction and decision making of AEEES based on ARRS. Refer to [17, 18] for the architecture and working process of ARRSs.

Figure 1 shows an architecture of ARRS-based AEEES, which includes the following components: elevator system, sensors, filter, predictor, decision maker, enactor, and public address system. *Elevator system* includes the elevator equipment, hoistway, machine room, and an traditional elevator control system. In AEEES, when emergency happens, the elevator control system will disable the regular dispatch mode, and receives instructions from enactor. *Sensors* perceive the situation of the building (including elevator systems) and occupants. Emergency sensors (e.g., heat detectors, smoke detectors, flame detectors, gas sensors) can detect emergencies. The situation of occupants (e.g., number of occupants in each floor/area, whether there are disable or injured people) can be monitored by using such as RFID [10] or cameras [19]. *Filter* filters out the trivial sensory data and generates important and useful information for predictor and decision maker. *Predictor* receives current situation from the filter, then outputs predictions about future hazards. *Decision maker* receives predictions from the predictor and current situation from the filter, then outputs next actions. *Enactor* receives next actions from decision maker, then give instructions to elevator control system. *Public address*

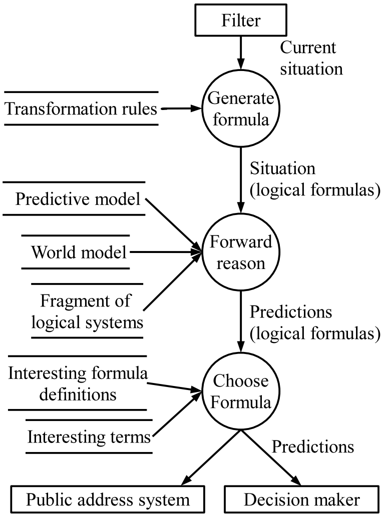


Fig. 2. Predictor's data flow diagram

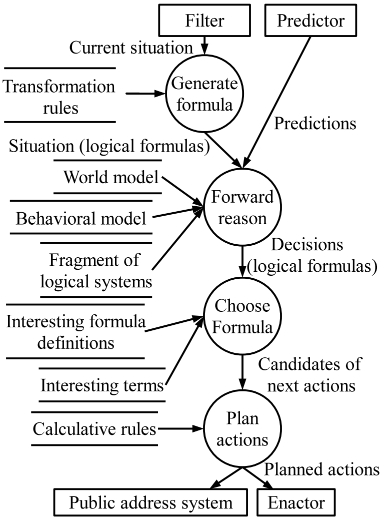


Fig. 3. Decision maker's data flow diagram

system (or other communication system such as using mobile phones [20]) gives occupants clear and understandable informative and instructive messages.

Figure 2 shows the data flow diagram of the predictor. First, formula generator transforms the filtered current situation to logical formulas. Second, the forward reasoning engine gets input of these logical formulas about current situation, the world model, the predictive model, and the fragment of logic system [5], then apply forward reasoning. Third, the formula chooser chooses predictions from all formulas reasoned out by forward reasoning engine, and transfers them to decision maker and other components such as public address system. In the ARRS, a *predictive model* is a set of empirical theories used for prediction, which are represented by logical formulas. The purpose of prediction of AEEES is to find out "what hazard will happen?" and "where the hazard will happen?" Furthermore, in order to predict more further, it is necessary to predict based on an existing prediction, i.e., "if we predict some hazards, then what hazards will happen after that?". In AEEES, the subjects of prediction are emergencies. Thus, there are different empirical theories for different emergencies. For example, in fire emergency, to predict fire spread, we have following empirical theorem "the upper floor of the afire floor will catch fire with high probability" [21, 22]: $\forall f_1 \forall f_2 (Upstairs(f_1, f_2) \land Afire(f_2) \land \neg Afire(f_1) \Rightarrow F(Afire(f_1)))$. To predict based on the prediction, we have: $\forall f_1 \forall f_2 (Upstairs(f_1, f_2) \land F(Afire(f_2)) \land \neg Afire(f_1) \Rightarrow U(Afire(f_2), F(Afire(f_1))))$, which means "if a floor will catch fire, then when this floor is really afire, its upstairs will catch fire". In this work, we use *temporal deontic relevant logics* [23] as the logic basis to both predict and make decisions, and the meanings of logical operators refer to [23].

Figure 3 shows the data flow diagram of the decision maker. There are two phases for making decisions: first phase is to make qualitative decision by forward reasoning, and second phase is to refine the decision by quantitative calculation. *Qualitative decision:* The progress of qualitative decision making by logic-based forward reasoning is similar with prediction by logic-based forward reasoning, where the main difference is to use behavioral model instead of predictive model. The purpose of qualitative decision is to find out "which actions should be taken?" and "which actions should be taken first?". Thus, the results of qualitative decision is a set of candidates of next actions with different priorities. In order to express the actions with priority in AEEES, we introduced priority constants (using floating number to express, smaller means higher priority): $CRITICAL>HIGH>MEDIUM>NORMAL>LOW> PLANNING$ and predicate $Pickup(f,p,et)$ that means an action to pick up people on floor f with priority p due to emergency type et. There are three types of actions in AEEES: reactive, anticipatory, and routine, which deal with ongoing hazards, predictive hazards, and full evacuation (to egress all occupants in the building) correspondingly. For example, in fire emergency, "the occupants of the fire floor are the occupants at the highest risk and should be the ones to be evacuated first by the elevators" [16], which can be expressed about reactive actions: $\forall f(Afire(f) \wedge Floor(f) \wedge Occupied(f) \Rightarrow O(Pickup(f, CRITICAL, FIRE)))$. For predictive hazards, we have: $\forall f(F(Afire(f)) \wedge Floor(f) \wedge Occupied(f) \Rightarrow O(Pickup(f, HIGH, FIRE)))$, and $\forall f \forall a(U(a, F(Afire(f))) \wedge Floor(f) \wedge Occupied(f) \Rightarrow O(Pickup(f, MEDIUM, FIRE)))$. For routine actions, down-peak are expressed as: $\forall et \forall f(FullEvacuation(et) \wedge Floor(f) \wedge Occupied(f) \Rightarrow P(Pickup(f, PriorCal(NORMAL, f), et)))$, which means when we want to apply full evacuation because of emergency type et, we can pick up each floor with priority $PriorCal(NORMAL, f)$. $PriorCal$ is a function to calculate priority, which result is $NORMAL - 0.001 \times FloorNumber(f)$. To consider full evacuation when fire emergency, can be expressed: $\forall r(Afire(r) \Rightarrow FullEvacuation(FIRE))$. Besides, because the aged and people with disabilities have higher priority in any emergency, we have: $\forall o \forall p \forall f(Person(o) \wedge Priority(o,p) \wedge Floor(f) \wedge In(o,f) \Rightarrow P(Pickup(f,p,ANY)))$. When certain emergency was eliminated (e.g. the fire dies out), the AEEES should know the evacuation should stop, thus we have: $\forall et(Eliminated(et) \Rightarrow P(StopEvacuation(et)))$. *Quantitative calculation:* In order to plan precise elevator car(s) dispatch, quantitative calculation is necessary. The decision maker maintains a priority set called $action\_container$, wich stores all candidates of actions according to their priority, i.e., to fetch the highest priority action when traversing. The decision maker also maintains a variable called $planned\_action$ for each elevator car. An algorithm to calculate quantitative $planned\_action$ is:

```
for each (elevator e1 in all elevators){
  if (e1 is free){
    for each (action in action_container){
      planRescuePeople = 0;
      for each (elevator e2 in all elevators)
        if (e2.planned_action.floor == action.floor)
          planRescuePeople += e2.CAPACITY;
      if(action.floor.peopleNumber+REDUNDANTNUM>planRescuePeople){
        e1.planned_action = action;
        break;
      }
    }
  }
}
```

The calculation is triggered by when an elevator car is free. When the elevator car arrives at the floor of its $planned\_action$, we set its $planned\_action$ as $NULL$.

6 Contrast Experiments

In order to evaluate AEEES, we took three groups of experiments to compare different EEESs: EEES using down-peak, reactive EEES (to rescue the floor with emergency first, then using down-peak), and AEEES.

The basic experiment scenario is that a building with 25 floors (almost all skyscrapers zone elevators with sky lobby, which is equivalent to superposition of several short building) and 8 elevators, uniform distribution of occupants, and the emergency is uncontrollable accelerated fire with random origin of fire. The experimental parameters' setting of fire emergency is based on [21, 22], and other parameters' setting is based on [24]. All occupants evacuate only by using elevators. We designed three groups of experiments: 1) origin of fire in different floor, 3,000 occupants in the building, 2) different fire spread speed, 3,000 occupants in the building, and 3) different number of occupants; while other parameters stay the same.

We consider the rescued ratio as the evaluation factor.

$$Rescued\,ratio = \frac{the\ number\ of\ rescued\ occupants}{the\ number\ of\ total\ occupants\ in\ the\ building} \times 100\%.$$

A good evacuation gets high rescued ratio.

Figure 4 shows the results of three groups of experiments. Using down-peak got lowest rescued ratio; AEEES got highest rescued ratio; and reactive EEES's rescued ratio was in between of them. To explain why such experimental results occur, we present a certain situation of first group of experiments. A fire emergency originated from 17 floor (F). Due to the fire alarm, all occupants began to evacuate from the building by using elevators. For down-peak, all 8 elevator cars were dispatched to 25F first. Because all occupants used elevators to evacuate, the car would stop at a floor which had occupants. However, the occupants threatened by the fire (might later die), such as occupants in 17F, cannot use the elevators in time. For reactive EEES, it sensed emergency on 17F, then dispatched all 8 cars to 17F first. Because there were 120 occupants in each floor and the maximum capacity of car is 20, these non-overloaded cars would stop at 16F and rescue occupants. (If a car is full, both reactive EEES and AEEES dispatch the car to evacuation floor without stopping.) After all occupants of 17F were rescued (the 18F was not afire at the moment), the reactive EEES adopted down-peak, thus all cars were dispatched to 25F. Later the 18F caught fire, reactive EEES sensed that and tried to rescue occupants on 18F. However, all cars were occupied at the moment, besides, the fire was severe, and there was little time left to rescue the occupants on 18F before they died. For AEEES, it filtered out the emergency on 17F and the situation of occupants, then transformed them into logical formulas (e.g., $Afire(17F)$, $Occupied(17F)$). Based on the predictive model in section 5 and the world model (e.g. $Floor(17F)$, $Floor(18F)$, $Upstairs(18F, 17F)$), AEEES deduced candidates of predictions. Then AEEES chose $F(Afire(18F))$ and $U(Afire(18F), F(Aire(19F))$, which means 18F would catch fire, and after that 19F would catch fire. For the qualitative decision, the AEEES chose $Pickup(17F, CRITICAL, FIRE)$, $Pickup(18F, HIGH, FIRE)$, $Pickup(19F, MEDIUM, FIRE)$, $Pickup(1F, PriorCal(NORMAL, 1F), FIRE))$, ..., and $Pickup(25F, PriorCal(NORMAL, 25F), FIRE))$, which means to rescue occupants on 17F with CRITICAL priority, 18F with HIGH priority, 19F with MEDIUM

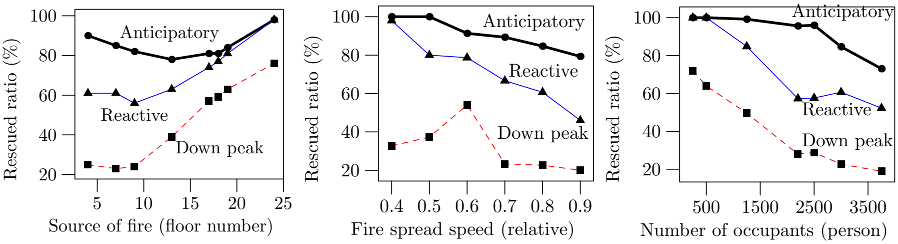


Fig. 4. Results of three groups of experiments

priority, and other floor with NORMAL priority besides higher floor has higher priority. Based on the calculation algorithm in section 5, the AEEES first dispatched 7 elevator cars to 17F and 1 elevator cars to 18F, then dispatched 6 cars to 18F and 2 cars to 19 floor, next dispatched 5 cars to 19F, after that dispatch 7 cars to 25F and 1 car to 24F.

We noticed sometimes anticipation becomes invalid when the emergency is too severe or there are too many occupants. Its essence is when carrying capacity of elevators is insufficient, AEEES can only rescue occupants in emergency floors, thus AEEES degenerates to reactive EEES.

7 Concluding Remarks

This paper presented anticipatory emergency elevator evacuation system (AEEES), its features, design, prototype, and evaluation. As a new type of emergency elevator evacuation systems (EEES), AEEESs show a new direction for EEESs.

We will consider the behavior of occupants, emergencies' accuracy, various emergencies, and evaluation from the viewpoint of efficiency in the future work. In fact, occupants may use stairs to evacuate, and they may spend a long time before they take evacuation [25], thus AEEESs should deal with this. One urgent issue is that our system do not consider the smoke of fire which is more deadly than flame. In order to construct good predictive model for AEEESs, we need qualitative experimental knowledge about emergencies. However, the current researches about emergencies are mainly numerical. It is a challenge to extract qualitative knowledge from these quantitative experimental knowledge.

References

[1] Bukowski, R.: International applications of elevators for fire service access and occupant egress in fires. CTBUH Journal (III), 28–584 (2010)

[2] CTBUH: Emergency evacuation elevator systems guideline. CTBUH (2004)

[3] Kuligowski, E., Bukowski, R.: Design of occupant egress systems for tall buildings. In: CIB World Building Congress 2004 (2004)

[4] Klote, J., Levin, B., Groner, N.: Feasibility of fire evacuation by elevators at FAA control towers. NISTIR 5445, NIST (1994)

[5] Cheng, J.: Temporal relevant logic as the logical basis of anticipatory reasoning-reacting systems. In: Proc. Computing Anticipatory Systems: CASYS - 6th International Conference, AIP Conference Proceedings. AIP, vol. 718, pp. 362–375 (2004)

[6] Barney, G.: Up-peak, down-peak & interfloor performance. Elevator World 47, 100–103 (1999)

[7] Klote, J., Levin, B., Groner, N.: Emergency elevator evacuation systems. In: Proc. 2nd Symposium on Elevators, Fire, and Accessibility, pp. 131–149. ASME (1995)

[8] Bukowski, R.: Addressing the needs of people using elevators for emergency evacuation. Fire Technology 48(1), 127–136 (2012)

[9] Levin, B., Groner, N.: Some control and communication considerations in designing an emergency elevator evacuation system. In: Proc. 2nd Symposium on Elevators, Fire, and Accessibility, pp. 190–193. ASME (1995)

[10] Luh, P., Xiong, B., Chang, S.: Group elevator scheduling with advance information for normal and emergency modes. IEEE Transactions on Automation Science and Engineering 5(2), 245–258 (2008)

[11] Kawai, K.: Fire control system for elevator, US Patent 7,413,059 (2008)

[12] Kawai, K.: Evacuation system and method for elevator control using number of people remaining, US Patent 7,637,354 (2009)

[13] Parrini, L., Spiess, P., Schuster, K., Finschi, L., Friedli, P., et al.: Method and system for emergency evacuation of building occupants and a method for modernization of an existing building with said system, US Patent 7,182,174 (2007)

[14] Bukowski, R.: Emergency egress strategies for buildings. In: Proc. 11th International Interflam Conference, pp. 159–168 (2007)

[15] Groner, N.: Selecting strategies for elevator evacuations. In: Proc. 2nd Symposium on Elevators, Fire, and Accessibility, pp. 186–189. ASME (1995)

[16] Proulx, G.: Evacuation by elevators: who goes first? In: Proc. Workshop on Use of Elevators in Fires and Other Emergencies, NRC Institute for Research in Construction, National Research Council Canada, pp. 1–13 (2004)

[17] Goto, Y., Kuboniwa, R., Cheng, J.: Development and maintenance environment for anticipatory reasoning-reacting systems. International Journal of Computing Anticipatory Systems 24, 61–72 (2011)

[18] Shang, F., Nara, S., Omi, T., Goto, Y., Cheng, J.: A prototype implementation of an anticipatory reasoning-reacting system. In: Proc. Computing Anticipatory Systems: CASYS - 7th International Conference, AIP Conference Proceedings. AIP, vol. 839, pp. 401–414 (2006)

[19] Kim, J., Moon, B.: Adaptive elevator group control with cameras. IEEE Transactions on Industrial Electronics 48(2), 377–382 (2001)

[20] Nakajima, Y., Shiina, H., Yamane, S., Ishida, T., Yamaki, H.: Disaster evacuation guide: Using a massively multiagent server and GPS mobile phones. In: Proc. 2007 International Symposium on Applications and the Internet, p. 2. IEEE (2007)

[21] Quintiere, J.G.: Fire growth: an overview. Fire Technology 33(1), 7–31 (1997)

[22] Morris, B., Jackman, L.: An examination of fire spread in multi-storey buildings via glazed curtain wall facades. Structural Engineer 81(9), 22–26 (2003)

[23] Cheng, J.: Temporal deontic relevant logic as the logical basis for decision making based on anticipatory reasoning. In: Proc. 2006 IEEE International Conference on Systems, Man and Cybernetics, vol. 2, pp. 1036–1041. IEEE (2006)

[24] Klote, J.: A method for calculation of elevator evacuation time. Journal of Fire Protection Engineering 5(3), 83–95 (1993)

[25] Proulx, G.: Evacuation time and movement in apartment buildings. Fire Safety Journal 24(3), 229–246 (1995)

A Stock Selective System by Using Hybrid Models of Classification

Shou-Hsiung Cheng

Department of Information Management, Chienkuo Technology University,
Changhua 500, Taiwan
shcheng@ctu.edu.tw

Abstract. Stock trade is a popular investing activity and during this activity, investors expect to gain higher profit with lower risk. Therefore, the problem of predicting stock returns has been an important issue for many years. This study is aimed on the discover relationship between financial data of public companies and return on investment by using data mining technology. The study propose a stock selective system by using hybrid models of classification. Use the hybrid models of association rules, cluster, and decision tree, it can provide meaningful decision rules for stock selection for intermediate- or long-term investors. Further, these rules are use to select some profitable stocks of the following years. The outcome evidences the higher return on investment in proposed model than general market average.

Keywords: Financial indexes, Relation rules, Cluster, Decision Tre**e.**

1 Introduction

Investment management plays an essential part in people's lives. Among many investment methods, the stock market is also one of that the most risky, but it can get high returns. Therefore, the problem of predicting stock returns has been an important issue for many years. Generally, there are two instruments to aid investors for doing prediction activities objectively and scientifically, which are technical analysis and fundamental analysis. Technical analysis examines historical data to identify correlations between volume and price that reflect the buying and selling behaviors of investors. However, fundamental analysis is a method of forecasting by way of using financial data to predict stock price. The fundamental analysis studies the causes stimulating changes in stock price, while the technical analysis studies the effect reflecting real situation about particular firms. Thus, the focus of this study is not on effects but on causes that should be seeking and exploring original sources actually. Therefore, selective a good stock is the first and the most important step for intermediate- or even long-term investment planning. In order to reduce risk, in Taiwan, the public stock market observation of permit period will disclose regularly and irregularly the financial statements of all listed companies. Therefore, this study employs data of fundamental analysis by using hybrid models of classification to

A. Selamat et al. (Eds.): ACIIDS 2013, Part I, LNAI 7802, pp. 127–134, 2013.

extract. Employing these meaningful decision rules, a useful stock selective system for intermediate- or long-term investors is proposed in this study.

Data mining is the processing of automatically searching large volumes of digital data for patterns by tools such as classification, cluster analysis, association rules and anomaly detection. Data mining is a more complex topic and can link with many fields like computer science and confluence of multiple disciplines such as information retrie- val, statistics, databases, algorithms, machine learning and pattern recognition. Clearly, data mining is a field of growing importance for increasing demand for AI, fast advance in IT techniques, and processing huge amount of digital data. In general, some related work considers a feature selection step to examine the usefulness of their chosen variables for effective stock prediction, e.g. [1]. This is because not all of features are informative or can provide high discrimination power. This can be called as the curse of dimensionality problem [2]. As a result, feature selection can be used to filter out redundant and/or irrelevant features from a chosen dataset resulting in more represen- tative features for better prediction performances [3]. The idea of combining multiple feature selection methods is derived from classifier ensembles [4]. The aim of classifier ensembles is to obtain high highly accurate ones. They are intended to improve the classification performance of a single classifier.The idea of combining multiple feature selection methods is derived from classifier ensembles (or multiple classifiers) [5]. The feature selection-based methods of continuous attributes are a valuable aspect of data mining. The classifier of hybrid models is intended to improve the classification performance of a single classifier. Therefore, the performance of hybrid models is likely better than one of the best single classifiers used in isolation.

The rest of the paper is organized as follows: In Section 2 an overview of the related works is introduced, while Section 3 presents the proposed procedure and briefly discusses its architecture. Section 4 describes analytically the experimental results. Finally, Section 5 shows conclusions of this paper.

2 Related Works

This study proposes a new stock selective system applying the association rules, cluster analysis and decision tree algorithm to verify that whether it can be helpful on prediction of the shares rose or fell for investors. Thus, this section mainly reviews related studies of the association rules, cluster analysis and decision tree.

2.1 Association Rules

Based on whether the ups and downs of each financial indicators and stock price then to calculate each financial indicator and stock price relationship, presented with various financial indicators and stock price to the growth of the number of times, the confidence of the various financial indicators and stock price can be calculated.

2.2 Cluster Analysis

Cluster analysis, also known as data cutting or non-supervised classification, it is a multivariate statistical analysis.The main purpose of the data records in the data collection, to be clustering into several clusters, so that the degree of similarity between a cluster of data records than other cluster record the degree of similarity. In this study, using the k-means cluster method, k-means cluster based on random numbers randomly selected group of center, that is, assuming that the data is divided into k groups, k number of randomly selected randomly as the k group the center closer to the group, and then to each number and the distance of the cluster centers, and moved to the period, and then calculate the new center value has been repeated to the elements within the cluster is no longer changes so far.

2.3 Decision Tree

ID3 decision tree algorithm is one of the earliest use, whose main core is to use a recursive form to cut training data. In each time generating node, some subsets of the training input tests will be drawn out to obtain the volume of information coming as a test. After selection, it will yield the greatest amount of value of information obtained as a branch node, selecting the next branch node in accordance with its recursively moves until the training data for each part of a classification fall into one category or meet a condition of satisfaction. C4.5 is the ID3 extension of the method which improved the ID3 excessive subset that contains only a small number of data issues, with handling continuous values-based property, noise processing, and having both pruning tree ability. C4.5 decision tree in each node use information obtained on the volume to select test attribute, to select the information obtained with the highest volume (or maximum entropy compression) of the property as the current test attribute node.

Let A be an attribute with k outcomes that partition the training set S into k subsets S_j (j = 1,..., k). Suppose there are m classes, denoted $C = \{c_1, \quad \cdots \quad , c_m\}$, and $p_i = \frac{n_i}{n}$ represents the proportion of instances in S belonging to class c_i, where $n = |S|$ and n_i is the number of instances in S belonging to c_i. The selection measure relative to data set S is defined by:

$$Info(S) = \sum_{i=1}^{m} p_i \log_2 p_i \tag{1}$$

The information measure after considering the partition of S obtained by taking into account the k outcomes of an attribute A is given by:

$$Info(S, A) = \sum_{j=1}^{k} \frac{|S_j|}{|S|} Info(S_i) \tag{2}$$

The information gain for an attribute A relative to the training set S is defined as follows:

$$Gain(S, A) = Info(S) - Info(S, A) \tag{3}$$

The Gain(S, A) is called attribute selection criterion. It computes the difference between the entropies before and after the partition, the largest difference corresponds to the best attribute. Information gain has the drawback to favour attributes with a large number of attribute values over those with a small number. To avoid this drawback, the information gain is replaced by a ratio called gain ratio:

$$GR(S, A) = \frac{\mathrm{Gain(S, A)}}{-\sum_{j=1}^{k} \frac{|\mathrm{S_i}|}{|S|} \log_2 \frac{|S_i|}{|S|}} \tag{4}$$

Consequently, the largest gain ratio corresponds to the best attribute.

3 Methodology

In this study, using the rate of operating expenses, cash flow ratio, current ratio, quick ratio, operating costs, operating income, accounts receivable turnover ratio (times), payable accounts payable cash (days), the net rate of return (after tax)net turnover ratio, earnings per share, operating margin, net growth rate, rate of return of total assets and 14 financial ratios, the use of data mining technology on the sample of the study association rules, cluster analysis, decision tree analysis taxonomy induction a simple and easy-to-understand investment rules significantly simplify the complexity of investment rules. The goal of this paper is proposes a straightforward and efficient stock selective system to reduce the complexity of investment rules.

3.1 Flowchart of Research Procedure

The study proposes a new procedure for a stock selective system. Figure 1 illustrates the flowchart of research procedure in this study.

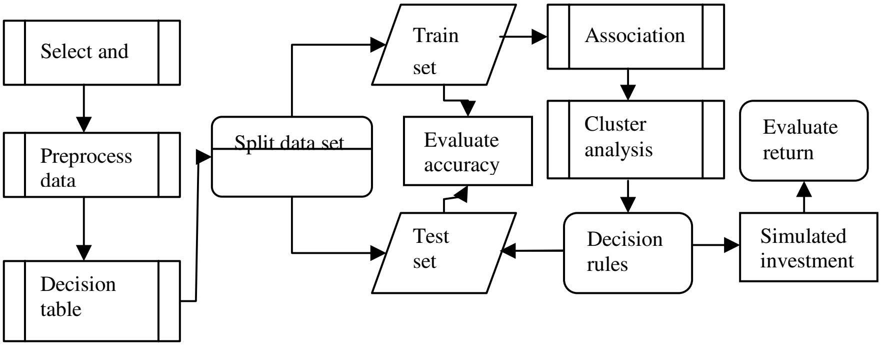


Fig. 1. Flowchart of research procedure

3.2 Flowchart of Research Procedure

This subsection further explains the proposed stock selective system and its algorithms. The proposed stock selective system can be devided into seven steps in detail, and its computing process is introduced systematically as follows:

Step 1: Select and collect the data.
Firstly, this study selects the target data that is collected from Taiwan stock trading system.

Step 2: Preprocess data.
To preprocess the dataset to make knowledge discovery easier is needed. Thus, firstly delete the records that include missing values or inaccurate values, eliminate the clearly irrelative attributes that will be more easily and effectively pro-cessed for extracting decision rules to select stock. The main jobs of this step includes data integration, data cleaning and data transformation.

Step 3: Build decision table.
The attribute sets of decision table can be divided into a condition attribute set and a decision attribute set. Use financial indicators as a condition attribute set, and whether the stock prices up or down as a decision attribute set.

Step 4: Association rule mining.
To calculate the confidence of various financial ratios and stock price change by using various financial ratios and price change association rules analysis. The confidence of various financial ratios and stock price change can be evaluated by equation (1)

$$C(X\rightarrow Y)=\frac{X\cap Y}{X}\text{，}(0<C\leq 1) \tag{1}$$

Step 5: Cluster analysis.
To cluster the condition attributes, the k-means cluster method through various confidence of the financial ratios is .

Step 6: Extract decision rules.
Based on condition attributes clustered in step 5 and a decision attribute (i.e. the stock prices up or down), generate decision rules by decision tree C5.0 algorithm.

Step 7: Evaluate and analyze the results and simulated investment.
For verification, the dataset is split into two sub-datasets: The 67% dataset is used as a training set, and the other 33% is used as a testing set. Furthermore, evaluate return of the simulated investment,

4 Empirical Analysis

4.1 The Computing Procedure

A practically collected dataset is used in this empirical case study to demonstrate the proposed procedure: The dataset for 993 general industrial firms listed in Taiwan stock trading system from 2008/01 to 2010/12 quarterly. The dataset contains 11916 instances which are characterized by the following 14 condition attributes: (i) operating expense ratio(A1), (ii) cash flow ratio(A2), (iii) current ratio(A3), (iv) quick ratio(A4), (v) operating costs(A5), (vi) operating profit(A6), (vii) accounts receivable turnover ratio(A7), (viii) the number of days to pay accounts(A8), (ix) return on equity (after tax) (A9), (x) net turnover(A10), (xi) earnings per share(A11), (xii) operating margin(A12), (xiii) net growth rate(A13), and (xiv) return on total assets growth rate(A14); all attributes are continuous data in this dataset. The computing process of the stock selective system can be expressed in detail as follows:

Step 1: Select and collect the data.

This study selects the target data that is collected from Taiwan stock trading system. Due to the different definitions of industry characteristics and accounting subjects, the general industrial stocks listed companies are considered as objects of the study. The experiment dataset contains 11916 instances which are characterized by 14 condition attributes and one decision attribute.

Step 2: Preprocess data.

Delete the 19 records (instances) that include missing values, and eliminate the 10 irrelative attributes. Accordingly, in total the data of 939 firms that consist of 14 attributes and 6793 instances are included in the dataset. The attributes information of the dataset is shown in Table 1.

Step 3: Build decision table.

The attribute sets of decision table can be divided into a condition attribute set and a decision attribute set. Use financial indicators as a condition attribute set, and whether the stock prices up or down as a decision attribute set.

Step 4: Association rule mining.

Use association rules to analyze various financial ratios and price change to calculate various financial ratios and Price Change reliability. Results of listed companies as shown in Table 1.

Table 1. The confidence of various condition attributes

Period	A1	A2	A3	A4	A5	A6	A7	A8	A9	A10	A11	A12	A13	A14
1	0.453	0.550	0.453	0.453	0.453	0.545	0.455	0.620	0.568	0.458	0.575	0.482	0.548	0.582
2	0.463	0.510	0.463	0.463	0.463	0.503	0.465	0.463	0.500	0.465	0.505	0.468	0.508	0.632
3	0.365	0.455	0.366	0.366	0.366	0.513	0.367	0.367	0.512	0.368	0.512	0.415	0.602	0.595
4	0.705	0.648	0.708	0.708	0.708	0.653	0.705	0.710	0.673	0.707	0.673	0.702	0.637	0.620
5	0.385	0.533	0.385	0.385	0.385	0.498	0.385	0.387	0.470	0.387	0.462	0.412	0.498	0.562
6	0.618	0.548	0.618	0.618	0.618	0.615	0.617	0.618	0.608	0.617	0.607	0.623	0.595	0.580
7	0.537	0.551	0.537	0.537	0.537	0.615	0.535	0.533	0.625	0.538	0.630	0.563	0.640	0.597
8	0.920	0.695	0.920	0.920	0.920	0.718	0.918	0.917	0.728	0.918	0.723	0.877	0.757	0.575
9	0.545	0.553	0.545	0.545	0.545	0.587	0.543	0.543	0.600	0.543	0.602	0.558	0.583	0.608
10	0.810	0.668	0.810	0.810	0.810	0.735	0.808	0.808	0.758	0.813	0.760	0.805	0.737	0.657
11	0.615	0.588	0.615	0.615	0.615	0.638	0.617	0.612	0.635	0.618	0.638	0.630	0.618	0.587
12	0.083	0.317	0.082	0.082	0.082	0.235	0.083	0.085	0.287	0.083	0.292	0.110	0.247	0.460
Average	0.542	0.552	0.542	0.542	0.542	0.571	0.542	0.555	0.580	0.543	0.582	0.554	0.581	0.588

Step 5: Cluster analysis.

To cluster the condition attributes, the k-means cluster method through various confidence of the financial ratios is. K-means algorithm, is through the average reliability of the data, with data similar to the way the data is divided into a number of clusters. By the reliability of the financial ratios of listed companies, k = 4 is divided into four groups for analysis in this study, clustering results as shown in Table 2.

Table 2. The Cluster of various condition attributes

No.	Condition attributes	Confidence	Cluster No.
1	operating expense ratio(A1)	0.542	1
2	cash flow ratio(A2)	0.552	4
3	current ratio(A3)	0.542	1
4	quick ratio(A4)	0.542	1
5	operating costs(A5)	0.542	1
6	operating profit(A6)	0.571	3
7	Accounts receivable turnover ratio(A7)	0.542	1
8	the number of days to pay accounts(A8)	0.555	4
9	return on equity (after tax) (A9)	0.580	2
10	net turnover(A10)	0.543	1
11	earnings per share(A11)	0.582	2
12	operating margin(A12)	0.554	4
13	net growth rate(A13)	0.581	2
14	return on total assets growth rate(A14)	0.588	**2**

Step 6: Extract decision rules.

Based on condition attributes clustered in step 5 and a decision attribute (i.e. the stock prices up or down), generate decision rules by decision tree C5.0 algorithm. C5.0 rule induction algorithm, is to build a decision tree recursive relationship between the interpretation of the field with the output field data divided into a subset of the and export decision tree rules, try a different part in the interpretation of the data with output field or the relationship of the results. The decission rule set of the financial ratios and the shares as shown in Table 3.

Table 3. The Decision rule set

No.	Decision rule set
1	If the total return on assets growth rate ≦ 0.07 and return on total assets growth rate ≦ -0.43 the shares fell.
2	If the return on total assets growth rate ≦ 0.07 and return on total assets growth rate> -0.43 and earnings per share of> 0.9 and return on total assets growth rate> -0.38, the shares fell.
3	If the return on total assets growth rate ≦ 0.07 and return on total assets growth rate> -0.43 and earnings per share of $ ≦ 0.9, the shares rose.
4	If the return on total assets growth rate ≦ 0.07 and return on total assets growth rate> -0.43 and earnings per share of> 0.9 and return on total assets growth rate ≦ -0.38 the shares rose.
5	If the return on total assets growth rate > 0.07, the shares rose.
6	If the operating margin ≦ 8.04, the shares fell.
7	If the operating margin > 8.04, the shares rose.

Step 7: Evaluate and analyze the results and simulated investment.

For verification, the dataset is split into two sub-datasets: The 67% dataset is used as a training set, and the other 33% is used as a testing set.

4.2 Simulated Investment

The rules generating from Table 4 get down on stock selection from the three quarters former listed companies rise in the rules in 2011 years. There are 227 in the first quarter in line with the rise in the rules, the average quarter rate of return of 22.68%; second quarter 56, the average quarter rate of return of 19.74%; 12 in the third quarter, the average quarter rate of return of 11.65%; 157 in the four quarter, the average quarter rate of return of 21.35%.

Table 4. The average quarter rate of return

	the average quarter rate of return	the broader market quarter rate of return
first quarter	22.68%	-0.02%
second quarter	19.74%	-0.01%
third quarter	11.65%	-0.21%
four quarter	21.35%	0.01%

5 Conclusion

The main contribution of this study is to use the association rules, cluster analysis, decision tree classification model data mining techniques to find rules which affect the company's share price change, from the information provided by the company's financial statements, then looking for an excellent company. From the above analysis, we can learn:

(1) The average rate of return on empirical result is much higher rate of return than the broader market.

(2) The three that impacts the ebb and flow of the listed company's share price are the growth rate of total return on assets, earnings per share and operating margin.

(3) Those companies the financial statements are beautiful can gain the investor recognition naturally, and the price naturally rises. So, to invest in stocks and to study the company's financial statements is a priority as well as a necessary work.

References

1. Abraham, A., Nath, B., Mahanti, P.K.: Hybrid Intelligent Systems for Stock Market Analysis. In: Alexandrov, V.N., Dongarra, J., Juliano, B.A., Renner, R.S., Tan, C.J.K. (eds.) ICCS 2001. LNCS, vol. 2074, pp. 337–345. Springer, Heidelberg (2001)
2. Huang, C.L., Tsai, C.Y.: A hybrid SOFM-SVR with a filter-based feature selection for stock market forecasting. Expert System with Applications 36(2), 1529–1539 (2009)
3. Chang, P.C., Liu, C.H.: A TSK type fuzzy rule based system for stock price prediction. Expert Systems with Application 34(1), 135–144 (2008)
4. Yu, L., Wang, S., Lai, K.K.: Mining Stock Market Tendency Using GA-Based Support Vector Machines. In: Deng, X., Ye, Y. (eds.) WINE 2005. LNCS, vol. 3828, pp. 336–345. Springer, Heidelberg (2005)
5. Kim, K.J.: Financial time series forecasting using support vector machines. Neurocomputing 55, 307–319 (2003)

An Efficient Method for Discovering Motifs in Large Time Series

Cao Duy Truong and Duong Tuan Anh

Faculty of Computer Science and Engineering,
Ho Chi Minh City University of Technology
caoduytruong@hcmunre.edu.vn, dtanh@cse.hcmut.edu.vn

Abstract. Time series motif is a previously unknown pattern appearing frequently in a time series. An efficient motif discovery algorithm for time series would be useful as a tool for summarizing massive time series databases as well as many other advanced time series data mining tasks. In this paper, we propose a new efficient algorithm, called EP-BIRCH, for finding motifs in large time series datasets. This algorithm is more efficient than MK algorithm and stable to the changes of input parameters and these parameters are easy to be determined through experiments. The instances of a discovered motif may be of different lengths and user does not have to predefine the length of the motif.

Keywords: time series motif, motif discovery algorithm, clustering algorithm, extreme points.

1 Introduction

A time series is a sequence of real numbers measured at equal intervals. Time series data arise in so many applications of various areas ranging from science, engineering, business, finance, economic, medicine to government. Nowadays, time series datasets in several applications becomes very large, with the scale of multi-terabytes and data mining task in time series data of such scale becomes very challenging.

Time series motifs are frequently occurring but previously unknown subsequences of a longer time series, which are very similar to each other. Since their formalization in 2002 by Lin et al. [4], several time series motif discovery algorithms have been proposed ([1], [4], [5], [6], [7], [12]). With most of these algorithms, user has to determine in advance the length of the motif and the distance threshold (range) for subsequence matching, which are the two parameters in most of the motif discovery algorithms. There have been a few algorithms proposed for finding time series motif with different lengths or variable lengths ([3], [9], [10]). However so far, to the best of our knowledge, there has been no time series motif discovery algorithm that can determine automatically the suitable length of the motif in a time series.

In this paper, we propose a new efficient algorithm for detecting time series motif which uses significant extreme points to determine motif candidates and then cluster motif candidates to find the most significant motif by using BIRCH algorithm. This

A. Selamat et al. (Eds.): ACIIDS 2013, Part I, LNAI 7802, pp. 135–145, 2013.

algorithm works directly in the original time series without any transformation for dimensionality reduction. The experiments on the real world datasets demonstrate that our proposed method outperforms the MK algorithm [6]. Besides, the proposed algorithm has three other major advantages. First, it can perform effectively with large time series datasets. Second, it is not sensitive to input parameters and these parameters can be determined easily. Third, the instances of a discovered motif may be of different lengths and user does not have to predefine the length of the motif.

2 Background

2.1 Time Series Motif and Bruce-Force Algorithm for Finding Motifs

Definition 1. *Time Series:* A time series $T = t_1,\ldots,t_N$ is an ordered set of N real-values measured at equal intervals.

Definition 2. *Similarity distance:* $D(s_1, s_2)$ is a positive value used to measure differences between two time series s_1, and s_2, relies on measure methods. If $D(s_1, s_2) < r$, where r is a real number (called *range*), then s_1 is similar to s_2 .

Definition 3. *Subsequence*: Given a time series T of length N, a subsequence C of T is a sampling of length $n < N$ of contiguous positions from T, that is, $C = t_p,\ldots,t_{p+n-1}$ for $1 \leq p \leq m - n + 1$.

Definition 4. *Time series motif:* Given a time series T, a subsequence C is called the most significant motif (or 1-motif) of T, if it has the highest count of the subsequences that are similar to it. All the subsequences that are similar to the motif are called *instances* of the motif.

Lin et al. ([4]) also introduced the brute-force algorithm to find the most significant motif (see Fig. 1). The input parameters of this algorithm are the length of the motif (n) and the distance threshold (r). Notice that the brute-force algorithm enumerates the non-trivial matches between subsequences. Given a time series T containing a subsequence C beginning at position p and a matching subsequence M beginning at q, we say that M is a *trivial match* to C if either $p = q$ or there does not exists a subsequence M' beginning at q' such that $D(C, M') > r$ and either $q<q'<p$ or $p<q'<q$. In the brute-force algorithm, we exclude the trivial matches so that only non-trivial matches are counted for detecting 1-motif in a time series.

This brute-force algorithm work directly on raw time series and requires $O(m^2)$ calls to the distance function (m is the length of the time series).

```
Algorithm Find-1-Motif-Brute-Force(T, n, r)
best_motif_count_so_far = 0
best_motif_location_so_far = null;
for i = 1 to length(T) – n + 1
   count = 0;  pointers = null;
   for j = 1 to length(T) – n + 1
```

```
      if Non_Trivial_Match ($C_{[i:\, i+n-1]}$, $C_{[j:\, j+n-1]}$, $r$) then
         count = count + 1;
         pointers = append (pointers, j);
       end
    end
    if count > best_motif_count_so_far then
      best_motif_count_so_far = count;
      best_motif_location_so_far = i;  motif_matches = pointers;
    end
   end
```

Fig. 1. The outline of brute-force algorithm for 1-motif discovery

2.2 Algorithms for Finding Motifs

Since the definition of time series motif was given in 2002 [4], several algorithms have been proposed to tackle the problem of time series motif discovery. The first algorithm that can find motifs in linear time is Random Projection, developed by Chiu et al. in 2003 [1]. It is an iterative approach and uses as base structure a collision matrix whose rows and columns are the SAX representation of each time series subsequence.

Mueen et al. in 2009 [6] proposed the first exact motif discovery algorithm, called MK algorithm, that works directly on raw time series data. One of the major disadvantages of the Random Projection algorithm and the MK algorithm is that they still execute very slowly with large time series data.

In this work, we improve the method for time series motif discovery proposed by Gruber et al. 2006 [2]. This method is based on the concepts of significant extreme points that was proposed by Pratt and Fink, 2002 [8]. The algorithm proposed by Gruber et al. for finding time series motifs consists of three steps: extracting significant extreme points, determining motif candidates from the extracted significant extreme points and clustering the motif candidates to determine the 1-motif through the cluster with the largest numbers of candidates. For convenience, in this paper, we call the algorithm proposed by Gruber et al. EP-C (Extreme Points and Clustering). When Gruber et al. proposed the EP-C algorithm, they apply it in signature verification and did not compare it to any previous time series motif discovery algorithms. Through our experiments done by Tin, 2012 [11], we found out that the EP-C is much more effective than Random Projection in terms of time efficiency and motif accuracy. But EP-C still needs some improvements.

2.3 Finding Time Series Motifs Using the MK Algorithm

Mueen et al. in 2009 [6] proposed the first exact motif discovery algorithm, called MK algorithm, that works directly on raw time series data. This algorithm uses the "nearest neighbor" definition of motif as follows. Time series motifs are pairs of subsequences which are very similar to each other.

Based on MK algorithm, we can modify it so that it can detect 1-motif in time series according the first formalized definition given by Lin et al. [4]. The modification can be done simply as follows: for each *i*-th subsequence of a longer time series, we use a linked list to store all the subsequences that match with the *i*-th subsequence. Later, the linked list with the largest number of matching subsequences will be the linked list associated with the 1-motif of the time series.

In the modified MK algorithm, we can also apply the three improvement techniques proposed by Mueen et al., 2009 ([6]). The three improvement techniques are (i) exploiting the symmetry of Euclidean distance, (ii) exploiting triangular inequality and reference point, and (iii) applying early abandoning.

Thank to the three techniques, the modified MK algorithm is up to three orders of magnitude faster than brute-force algorithm. More details about the three improvement techniques, interested reader can refer to [6]. In this work, we will use the modified MK algorithm as the baseline algorithm to compare with our proposed algorithm for finding time series motif.

2.4 Finding Significant Extreme Points

To extract a temporally ordered sequence of motif candidates, significant extreme points of a time series have to be found. The definition of significant extreme points, given by Pratt and Fink, 2002 [8] is as follows.

Definition 5. *Significant Extreme Points:* A univariate time series $T = t_1,...,t_N$ has a *significant minimum* at position m with $1 < m < N$, if $(t_i, . . . , t_j)$ with $1 \leq i < j \leq N$ in T exists, such that t_m is the minimum of all points of this subsequence and $t_i \geq R\times t_m$, $t_j \geq R\times t_m$ with user-defined $R \geq 1$.
Similarly, a *significant maximum* is existent at position m with $1 < m < N$, if a subsequence $(t_i, . . . , t_j)$ with $1 \leq i < j \leq N$ in T exists, such that t_m is the maximum of all points of this subsequence and $t_i \leq R\times t_m$, $t_j \leq R\times t_m$ with user-defined $R \geq 1$.

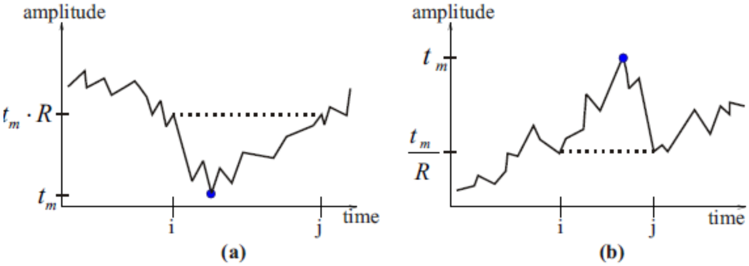


Fig. 2. Illustration of Significant Extreme Points: (a) Minimum, (b) Maximum

Notice that in the above definition, the parameter R is called *compression rate* which is greater than one and an increase of R leads to selection of fewer significant extreme points. Fig. 2 illustrates the definition of significant minima (a) and maxima (b). Given a time series *T*, starting at the beginning of the time series, all significant minima and maxima of the time series are computed by using the algorithm given in [8].

The significant extreme points can be the starting point or ending point of a motif instances. Basing on the extracted significant points we can extract the motif candidates from a time series and then cluster them using BIRCH algorithm.

2.5 BIRCH Clustering

BIRCH is designed for clustering a large amount of numerical data by integration of hierarchical clustering at the initial stage and other clustering methods, such as iterative partitioning at the later stage ([13]). It introduces two main concepts, *clustering feature* and *clustering feature tree* (CF tree), which are used to summarize cluster representations. These structures help the clustering method achieve good speed and scalability in large databases. BIRCH is also effective for incremental and dynamic clustering of incoming objects.

Given *N* *d*-dimensional points or objects $\overrightarrow{x_i}$ in a cluster, we can define the centroid $\overrightarrow{x_0}$, the radius *R*, and the diameter *D* of the cluster as follows:

$$\overrightarrow{x_0} = \frac{\sum_{i=1}^{N} \overrightarrow{x_i}}{N}$$

$$R = \left(\frac{\sum_{i=1}^{N} (\vec{x}_i - \vec{x}_0)^2}{N} \right)^{\frac{1}{2}}$$

$$D = \left(\frac{\sum_{i=1}^{N} \sum_{j=1}^{N} (\vec{x}_i - \vec{x}_j)^2}{N(N-1)} \right)^{\frac{1}{2}}$$

where *R* is the average distance from member objects to the centroid, and *D* is the average pairwise distance within a cluster. Both *R* and *D* reflect the tightness of the cluster around the centroid. A clustering feature (CF) is a triplet summarizing information about clusters of objects. Given *N* *d*-dimensional points or objects in a subcluster, then the CF of the cluster is defined as

$$\mathrm{CF} = (N, \overrightarrow{LS}, SS)$$

where *N* is the number of points in the subcluster, $\overrightarrow{LS}$ is the linear sum on *N* points and *SS* is the square sum of data points.

$$\overrightarrow{LS} = \sum_{i=1}^{N} \vec{x}_i$$

$$SS = \sum_{i=1}^{N} \vec{x}_i^{\,2}$$

A clustering feature is essentially a summary of the statistics for the given subcluster: the zero-th, first, and second moments of the subcluster from a statistical point of view. Clustering features are *additive*. For example, suppose that we have two disjoint clusters, C_1 and C_2, having the clustering features, CF_1 and CF_2, respectively. The clustering feature for the cluster that is formed by merging C_1 and C_2 is simply $CF_1 + CF_2$. Clustering features are sufficient for calculating all of the measurements that are needed for making clustering decisions in BIRCH.

A CF tree is a height-balanced tree that stores the clustering features for a hierarchical clustering. By definition, a nonterminal node in the tree has descendents or "children". The nonleaf nodes store sums of the CFs of their children, and thus summarize clustering information about their children. Each entry in a leaf node is not a single data objects but a subcluster. A CF tree has two parameters: branching factor (B for nonleaf node and L for leaf node) and threshold T. The branching factor specifies the maximum number of children in each nonleaf or leaf node. The threshold parameter specifies the maximum diameter of the subcluster stored at the leaf nodes of the tree. The two parameters influence the size of the resulting tree.

BIRCH applies a multiphase clustering technique: a single scan of the data set yield a basic good clustering, and one or more additional scans can (optionally) be used to further improve the quality. The BIRCH algorithm consists of four phases as follows.

Phase 1: (*Building CF tree*) BIRCH scans the database to build an initial in-memory CF tree, which can be view as a multilevel compression of the data that tries to preserve the inherent clustering structure of the data.

Phase 2: [optional] (*Condense data*) Condense into desirable range by building a smaller CF tree.

Phase 3: (*Global Clustering*) BIRCH applies a selected clustering algorithm to cluster the leaf nodes of the CF tree. The selected algorithm is adapted to work with a set of subclusters, rather than to work with a set of data points.

Phase 4: [optional] *Cluster refining*

After the CF tree is built, any clustering algorithm, such as a typical partitioning algorithm, can be used in Phase 3 with the CF tree built in the previous phase. Phase 4 uses the centroids of the clusters produced by Phase 3 as seeds and redistributes the data points to its closest seed to obtain a set of new clusters.

3 The Proposed Method – Combination of Significant Extreme Points and BIRCH

The proposed method, called EP-BIRCH (Extreme points and BIRCH clustering), is an improvement of the EP-C described in Section 2. The EP-C algorithm by Gruber et al. [2] uses hierarchical agglomerative clustering (HAC) algorithm for clustering which is not suitable to large scale time series datasets. In our proposed method, we use BIRCH algorithm to cluster motif candidates rather than using HAC algorithm. BIRCH is especially suitable for clustering very large time series datasets. Besides, in the EP-C algorithm, each motif candidate is determined by three contiguous extreme points, but in our proposed method, motif candidate is determined by *n* contiguous extreme points where *n* is selected by user.

EP-BIRCH consists of the following steps:

Step 1: We extract all significant extreme point of the time series *T*. The result of this step is a sequence of extreme points $EP = (ep_1, \ldots, ep_l)$

Step 2: We compute all the motif candidates iteratively. A motif candidate $MC_i(T)$, $i = 1, \ldots, l - 2$ is the subsequence of *T* that is bounded by the *n* extreme points ep_i and ep_{i+n-1}. Motif candidates are the subsequences that may have different lengths.

Step 3: Motif candidates are the subsequences that may have different lengths. To enable the computation of distances between them, we can bring them to the same length using homothetic transformation. The same length here is the average length of all motif candidates extracted in Step 2.

Step 4: We build the CF tree with parameters *B* and *T*. We insert to the CF tree all the motif candidates found in Step 3. We apply k-Means as Phase 3 of BIRCH to cluster the leaf nodes of the CF tree where *k* is equal to the number of the leaf nodes in the CF tree.

Step 5: Finally we find the subcluster in the CF tree with the largest number of objects. The 1-motif will be represented by that cluster.

In the Step 3, to improve the effectiveness of our proposed method, we apply *homothety* for transforming the motif candidates with different lengths to those of the same length rather than spline interpolation as suggested in [2]. Spline interpolation is not only complicated in computation, but also can modify undesirably the shapes of the motif candidates. Homothety is a simpler and more effective technique which also can transform the subsequences with different lengths to those of the same length.

Homothety is a transformation in affine space. Given a point O and a value $k \neq 0$. A homothety with center O and ratio k transforms M to M' such that $\overrightarrow{OM'} = k \times \overrightarrow{OM}$. Fig. 3. shows a homothety with center *O* and ratio k = ½ which transforms the triangle MNP to the triangle M'N'P'.

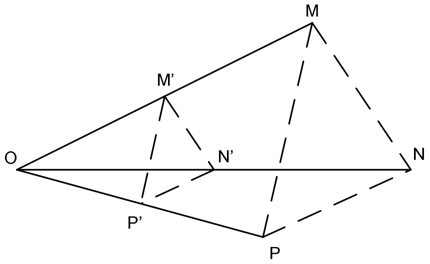


Fig. 3. Homothetic Transformation

Homothety can preserve the shapes of any curves under the transformation. Therefore, it can be used to align a longer motif candidate to a shorter one. The algorithm that performs homothety to transform a motif candidate *T* with length ***N*** (*T* = {Y_1,…,$\mathrm{Y_N}$}) to motif candidate of length *N'* is given as follows.

1. Let *Y_Max* = Max{Y_1,…,$\mathrm{Y_N}$}; *Y_Min* = Min {Y_1,…,$\mathrm{Y_N}$}
2. Find a center *I* of the homothety with the coordinate: *X_Center* = *N*/2, *Y_Center* = (*Y_Max* + Y_Min)/2
3. Perform the homothety with center *I* and ratio *k* = *N'*/*N*.

Notice that in Step 4 of our proposed method, if the parameters *B* and *T* are well selected, the number of the leaf nodes in the CF tree is approximately the suitable number of the clusters for the particular set of motif candidates.

4 Experimental Evaluation

In this experiment, we compare our EP-BIRCH algorithm to the modified MK algorithm described in Section 2. The MK algorithm is selected for comparison since it is the most recent proposed motif discovery algorithm which has remarkable efficiency. We implemented the two motif discovery algorithms with Microsoft Visual C# and conducted the experiment on a Core i7, Ram 4GB PC. We tested the algorithms on six publicly available datasets. The datasets are described as follows.

1. Monthly air temperatures in Tokyo, measured at the Station No:47662, from 01/1876 to 06/2012[1]
2. Natural Gas Futures Contract 1 (Dollars per Million BTU) from 31/12/1993 to 13.07.2012[2].
3. Power Demand by ECN, displayed as a function of hours and days[3]
4. Euro/US Dollar Exchange rates from 28.03.2005 to 28.03.2006, measured at every 5 minutes[4]
5. Koski ECG (electrocardiogram) dataset[5]
6. Sea level dataset, measured at Coastal Ocean Observation Network TCOON, at every 6 minutes[6]

The comparison is in terms of running time and efficiency. Here we evaluate the efficiency of each algorithm by simply considering the ratio of how many times the Euclidean distance function must be called by this algorithm over the number of times it must be called by the brute-force algorithm given in Section 2. The efficiency value

[1] http://www.data.jma.go.jp/obd/stats/etrn/view/monthly_s3_en.php?block_no=47662&view=7

[2] http://www.eia.gov/dnav/ng/hist/rngc1d.htm

[3] http://www.cs.ucr.edu/~eamonn/Keogh_Time_Series_CDrom

[4] http://www.forexpros.com/currencies/eur-usd-historical-data

[5] http://www.cs.ucr.edu/~eamonn/iSAX/koski_ecg.dat

[6] http://lighthouse.tamucc.edu/pq

is always less than 1; the method with lower efficiency value is better. The experimental results for the efficiency of the two motif discovery algorithms, EP-BIRCH and MK, on the six datasets are shown in Table 1.

Table 1. Experimental results on the two algorithms over 6 datasets

Data	**Length**	**Motif average length**	**Efficiency (%)**		**Runtime (sec)**	
			MK	**EP-BIRCH**	**MK**	**EP-BIRCH**
Tokyo air tempetature	1639	13	9.16	1.537	5.807	0.382
Natural Gas	4638	34	33.64	5.569	12.789	1.002
Power	35040	99	2.43	0.112	120.741	3.015
Euro/USD	78893	36	0.7	0.13	532.023	13.383
ECG	144404	157	6.15	0.004	9146.328	4.344
TCOON	175200	112	Out of memory	0.046	Out of memory	21.167

From the experimental results in Table 1 we can see that:

1. EP-BIRCH is more efficient than MK in terms of CPU times and efficiency values. EP-BIRCH is up to four orders of magnitude faster than brute-force algorithm.
2. EP-BIRCH can find motifs on large time series dataset. With large datasets such as TCOON (175200 data points), MK can not work, while EP-BIRCH can find the motif in a very short time (21 seconds)
3. EP-BIRCH can find motif instances with different lengths.
4. The performance of EP-BIRCH is quite stable when some input parameters are changed.
5. When the input parameters are set with suitable values, the 1-motif found by EP-BIRCH is exactly similar to the 1-motif found by the bruce-force algorithm.

The Effects of Parameters on the Performance of EP-BIRCH

EP-BIRCH requires from user 4 parameters: *R* (compression rate for computing the significant extreme points, *n* (the number of significant extreme points for each motif candidate, *B* (the branching factor of a nonterminal node in CF tree) and *T* (the maximum diameter of the subclusters stored in the leaf nodes in CF tree). The length of motif candidates is determined by the two parameters *R* and *n*. With larger *R*, less extreme points are extracted and the distance between two extreme points will become larger. With smaller *R*, more extreme points are extracted and the distance between two extreme points will become shorter. However when we increase *n* we will obtain motif candidates with larger length. Therefore, we can determine easily the values of *R* and *n* such that we obtain the desirable motif length.

We conducted an experiment to compare the motifs detected by EP-BIRCH when we change the values of parameters T and B. Experiments on the ECG dataset shows the following results:

- For n = 2: When T changes from 0.4 to 1.4 and for all different values of B, all the detected motifs are the same.
- For n = 3: When T changes from 1.0 to 2.0 and for all different values of B, all the detected motifs are the same.

The experimental results reveal that EP-BIRCH is quite stable when the two parameters T and B change in some given ranges.

5 Conclusions

We have introduced a new method for discovering motifs in time series which can work efficiently on large time series datasets. This method, called EP-BIRCH, is based on extracting significant extreme points and clustering the motif candidates by using BIRCH algorithm. The experiments on the real world datasets demonstrate that our proposed method outperforms the MK algorithm in terms of efficiency. Notice that our proposed method requires only one single scan over the entire time series dataset. Therefore, we can apply EP-BIRCH not only for discovering motifs on large time series datasets, but also for finding motifs in streaming time series.

As for future work, we plan to extend the proposed method in finding motifs in streaming time series and create a disk-aware version of our algorithm to allow the exploration of truly massive time series datasets.

References

1. Chiu, B., Keogh, E., Lonardi, S.: Probabilistic discovery of time series motifs. In: Proc. of 9th Int. Conf. on Knowledge Discovery and Data Mining (KDD 2003), pp. 493–498 (2003)
2. Gruber, C., Coduro, M., Sick, B.: Signature verification with dynamic RBF network and time series motifs. In: Proc. of 10th International Workshop on Frontiers in Hand Writing Recognition (2006)
3. Li, Y., Lin, J.: Approximate Variable-Length Time Series Motif Discovery Using Grammar Inference. In: Proceedings of the Tenth International Workshop on Multimedia Data Mining, Washington, D.C (July 25, 2010)
4. Lin, J., Keogh, E., Patel, P. and Lonardi, S.: Finding Motifs in Time Series. In: Proceedings of the 2nd Workshop on Temporal Data Mining, at the 8th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (2002)
5. Liu, Z., Yu, J.X., Lin, X., Lu, H., Wang, W.: Locating Motifs in Time-Series Data. In: Ho, T.-B., Cheung, D., Liu, H. (eds.) PAKDD 2005. LNCS (LNAI), vol. 3518, pp. 343–353. Springer, Heidelberg (2005)
6. Mueen, A., Keogh, E., Zhu, Q., Cash, S., Westover, B.: Exact Discovery of Time Series Motif. In: Proc. of 2009 SIAM International Conference on Data Mining, pp. 1–12 (2009)

7. Patel, P., Keogh, E., Lin, J., Lonardi, S.: Mining motifs in massive time series databases. In: Proc. of IEEE Int. Conf. on Data Mining, pp. 370–377 (2002)
8. Pratt, K.B., Fink, E.: Search for patterns in compressed time series. International Journal of Image and Graphics 2(1), 89–106 (2002)
9. Tang, H., Liao, S.: Discovering Original Motifs with Different Lengths from Time Series. Knowledge-based Systems 21(7), 666–671 (2008)
10. Tanaka, Y., Iwamoto, K., Uehara, K.: Discovery of Time-Series Motif from Multi-Dimensional Data Based on MDL Principle. Machine Learning 58(2-3), 269–300 (2005)
11. Tin, H.: N: Time Series Motif Discovery based on Important Extreme Points, Master Thesis, Faculty of Computer Science and Engineering, Ho Chi Minh University of Technology, Vietnam (July 2012)
12. Yankov, D., Keogh, E., Medina, J., Chiu, B., Zordan, V.: Detecting Time Series Motifs Under Uniform Scaling. In: Proc. of 13th ACM SIGKDD Int. Conf. on Knowledge Discovery and Data Mining (KDD 2007), pp. 844–853 (2007)
13. Zhang, T., Ramakrishnan, R., Livny, M.: BIRCH: An efficient data clustering method for very large databases. SIGMOD Rec. 25(2), 103–114 (1996)

CFTL – Flash Translation Layer for Column Oriented Databases

Krzysztof Kwiatkowski and Wojciech Macyna

Institute of Mathematics and Computer Science
Wrocław University of Technology
Poland
krzkw87@gmail.com, wojciech.macyna@pwr.wroc.pl

Abstract. The flash memory becomes a very popular storage device. Recently, it has replaced the traditional hard disc due to its shock - resistance and low power consumption. However, the flash memory architecture has some limitations related to the data management. Data cannot be updated in-place. That is why the flash translation layer (FTL) must be used. Nowadays, the flash memory seems to be the mature enough storage technology to work with databases. Among many database models, the column oriented databases have attracted the attention. This approach indicates the fact that data from the columns are packed together in the memory blocks. In this paper, we propose the Column Flash Translation Layer (CFTL) as an effective tool to store the data of the column oriented database in blocks of the flash memory.

Keywords: flash memory, flash translation layer, column oriented database.

1 Introduction

The flash memory is a very popular storage device. This is due to its shock-resistance, power economy and non-volatile nature. As an opposite to the traditional hard discs, it doesn't contain any moving parts inside. So it can be used in mobile applications such as: PDA devices, sensor networks, cell phones, music players and so on. Recently, the popularity of the flash storage has increased because of its capacity and reliability. The flash memory has different architecture than a hard disc. It consists of blocks and every block contains the fixed number of pages (32 or 64). One flash memory page has typically 512-2048 bytes. The read and write operations are performed at the page granularity, but the way of data updating is completely different. The page cannot be overwritten. If the data modification occurs, the new version must be written to a free page. Simultaneously, the old page is checked as "obsolete" and will be reclaimed later. To remove data physically, the whole block must be erased. As a consequence, the other valid pages must be replaced to the other block of the flash memory. The above described situation requires the use of a special software called the Flash Translation Layer (FTL) [1]. The FTL maps the logical address to the physical address in the flash memory. There are several types of the FTL which are based on the different mapping granularity. In the block-level translation layer, the logical block is mapped to the physical block. In this case, when one page of the block is changed, all other pages of

A. Selamat et al. (Eds.): ACIIDS 2013, Part I, LNAI 7802, pp. 146–155, 2013.

this block must be rewritten to the new location. It imposes the overhead that makes this kind of the FTL ineffective. However, it has been noticed that not the block-level translation layer but the page-level translation layer is more popular these days. It maps particular logical pages to the physical pages in the flash memory. For example, if the application accesses the data on the logical page 50, the FTL redirects it to the physical page 100. Then, if the data are updated, they are rewritten to the physical page 101 and the page-level mapping is updated. In this way, the application requests the logical page 50 and it is not aware that the physical location of the data has changed. However, this technique cannot be acceptable for the devices with limited memory because the page mapping structure may be too big for their resources. In order to overcome those limitations many hybrid-level FTLs have been proposed. Such approaches assume that all the physical blocks are separated between data blocks (D-blocks) and update blocks (U-blocks). The D-blocks are mapped on the block level and the U-blocks on the page level. When the data in a D-block change, they are written to the other location in an U-block and the page of the D-block is obsoleted. If there is no empty space in the U-blocks, a garbage collector moves the valid pages to the new D-blocks, erases the blocks containing the obsolete pages and creates the new U-blocks for the further use [1].

Another important features in this field are related to the time delay and energy consumption. The random reads in the flash memory are faster than in a magnetic disc. On the other hand, the random writes to the flash memory are much slower than reads. Moreover, the write operation is much more energy consuming than the read one.

One of the research directions considers the application of the flash memory in databases. Most vendors of the database systems use row oriented architecture. In this model, the records are placed contiguously in a storage. To retrieve the specified attributes of the row, all fields of the record must be pushed out to the disc. Such architecture is write-optimal, what in consequence makes it more appropriate for the OLTP-style applications.

An alternative approach to the above-mentioned one, is the column oriented database model. It is particularly suitable for data warehouses and other read optimized systems. The key idea is to store the data column-wise. In that case, each memory block contains the data of the same attribute. The biggest advantage of this architecture is that the queries don't need to retrieve all data of the required rows, but only the values of the specified attributes. Lately, several prototypes have been implemented. In the decomposition storage model [2] each column of the table is stored as a pair: (tuple id, attribute values). MonetDB/X100 [3] operates on the columns as vectors in memory and C-Store presented in [4] does not explicitly store the tuple identifier associated with the column.

In this paper, we propose an effective storage method for the column oriented databases in the flash memory architecture. We describe the method of a column oriented storage, take into account the flash memory constraints and overcome such limitations using Column Flash Translation Layer (CFTL). In our approach, data are stored in the entry sequence. To retrieve the row from the particular attributes, we don't store the row identifiers, but simply calculate their location in the column using our CFTL. Moreover, we present the data modification techniques and the precise data reclamation procedure. We conduct several experiments which confirm the effectiveness of our approach empirically. The paper focuses only on the storage method in the flash memory environment. Another aspects of the database systems are out of the scope of this paper.

Our paper is organised as follows: In section 2, we describe our approach. In section 3, we present calculations and after that we describe the experiments. In the last section, we summarize the paper.

2 CFTL

In this section, we present our storage method and the Column Flash Translation Layer (CFTL) to facilitate the data management for the column oriented database. The column oriented database supports the standard relational logical model. The database consists of a collection of tables and those tables contain attributes. For every attribute of the table, we define a collection of blocks and flash memory pages, in which the data for this attribute reside. In this approach, data are stored column-wise in flash memory blocks.

The CFTL is divided into two structures: CFTL_META and CFTL_DATA. In CFTL_META we store the meta data (see figure 1) which includes the following components: ID_Tab, ID_Attr, $SBlock\_i$, $Attr\_Name$ and $Attr\_Size$. They denote a table identifier, an attribute identifier, the first block in the S-Map, the attribute name and its size, respectively. As far as the CFTL_DATA is concerned, it consists of two mapping structures: S-Map and P-Map. Both structures facilitate traversing through the data of the particular column.

The first structure contains an ordered collection of D-blocks while the second one comprises of a collection of flash memory pages from various U-blocks. To one attribute from the database table may be assigned only one S-Map structure. With each D-block of the S-Map may be associated at most one P-Map structure. The first entry of the P-Map is the identifier of the D-block. The successive entries are the pairs of: block number ($PBlock\_i$), page number ($PPage\_i$), which points to the page $PPage\_i$ in the block $PBlock\_i$. All the dependencies are shown on figure 1.

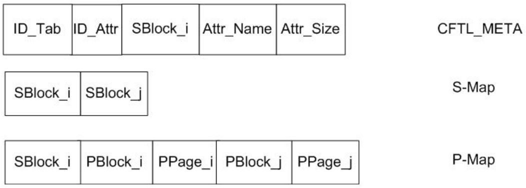


Fig. 1. CFTL_DATA

Supposing that we store the table EMP consisting of 4 attributes in our database (see the figure 1). The figure 2 shows an example of the data binding. On the top of the figure, there are pages containing the meta data. One S-Map is connected with each attribute. The data of the attribute $Dept$ are stored in two pages: (0,0) and (0,1), which belong

Table 1. Relation EMP

Name	Age	Dept	Salary
Bob	30	Developer	3 000
Tom	33	Electrician	5 000
Andrew	27	Architect	7 000

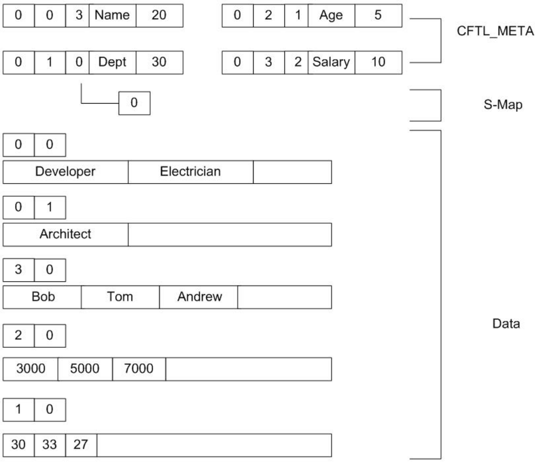


Fig. 2. Data binding

to the block 0. Thus, this example demonstrates that every S-Map structure consists of one entry because all the data for each attribute are stored in one memory block. If the data have not been modified yet, the P-Map is not needed.

2.1 Operations on Data

In this section, we describe the methods of performing such operations as: $insert$, $update$, $delete$. Besides, we point to the data retrieving issues. These methods utilize the main memory buffer before the processing. First, the data are temporarily held in the buffer. When it is full, the data are written to the free page of the flash memory. The main purpose of using the buffer is to collect data and write them in one batch to the flash memory. In the warehouse environment, the data are appended to the database. CFTL should preserve the timestamp ordering of the inserted records.

Insert. Supposing that we insert three tuples to the table EMP: ('John', 'Clerk', 28, 1550), ('Tomas', 'Accounter', 34, 2000), ('Mary', 'Secretary', 22, 1100). First, they are stored in the main memory buffer. Then, they are packed column-wise to the new flash memory pages. In this case, data are allocated in four flash memory pages: $Clerk$ and $Accounter$ are written to page $(0, 2)$, $Secretary$ to $(0, 3)$, $John, Thomas, Mary$ to page $(3, 1)$. The new values of attributes: $Salary$ and Age are stored in pages: $(2, 1)$

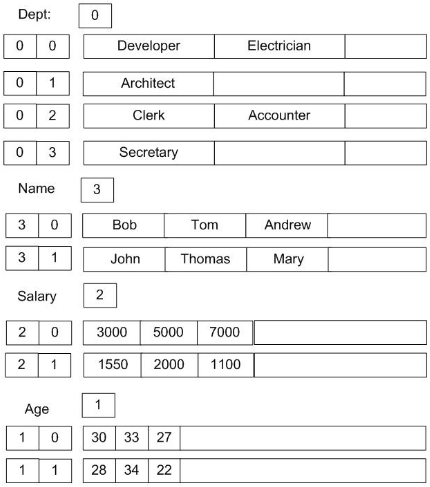


Fig. 3. Insert operation

and $(1, 1)$, respectively (see the figure 3). Please note that there are no changes in the CFTL_DATA structure because the inserted data are loaded to the same block of the flash memory.

Delete. To delete a row from the table, all the pages containing its data must be retrieved to the main memory. Afterwards the deleted values are removed and the other data of the pages must be rewritten to the other location of the flash memory.

Supposing that two rows are deleted from the table T which has three attributes: A, B, C (see the figure 4). Every attribute has the different size so the various number of data can be packed in one page of the flash memory. The figure 4 presents the data and the CFTL_DATA structure of T before (left side) and after (right side) performing the delete operation which is described hereinafter.

If the data of T have not been changed before, the S-Map structures are sufficient to reflect the state of the table. Each of the S-Maps contains two block numbers, as it is shown in the left lower part of the figure 4.

When the rows $R1$ and $R2$ are supposed to be deleted, three memory pages $(1, 2)$, $(2, 1)$ and $(3, 1)$, will be fetched to the main memory. After that, the delete operation is processed and the other data are dumped back to the free pages of the flash memory. Now the CFTL_DATA structure is updated. Pages $(1, 2)$ and $(3, 1)$ are replaced by $(8, 1)$ and $(9, 3)$, respectively. Moreover, pages $(1, 2)$, $(2, 1)$ and $(3, 1)$ are invalidated. If the D-block contains the ”obsolete” pages, the P-Map will be needed. The right side of the figure 4 shows the P-Map structure for the blocks: 1, 2, and 3. For example, the third entry of the P-Map connected with block 1 is the page $(8, 1)$. This is because the page $(1, 2)$ was replaced by $(8, 1)$.

Update. The updating technique is similar to the performing of the delete operation. The page, which contains the data, that will be updated, is fetched to the main memory

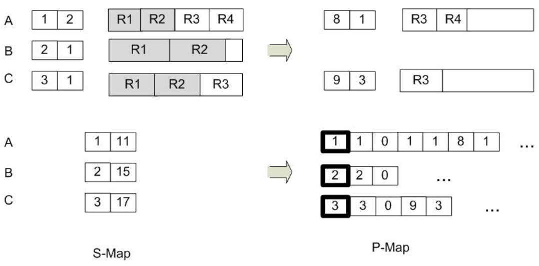


Fig. 4. Delete operation

and after committing the update process is written back to one of the U-block pages. The P-Map structure must be changed thereafter.

Data Retrieving. The main disadvantage of the column oriented databases is that the columns must be stitched back together in order to create a row. It makes the scan queries complicated. Our approach resolves this problem in a similar manner to [4]. We don't store the row identifier in the flash memory. Instead, we calculate the appropriate positions of data using the S-Map and P-Map structure. In that way, we are able to derive the data from all the queried columns and create output rows.

2.2 Data Compacting

The very intense workload may cause the data scattered over many blocks, yet it may create many invalid pages. These pages must be reclaimed for further usage and the data should be compacted in the new D-blocks. The reclamation procedure is invoked, if there is no space in the U-blocks or there are too many invalid pages in the D-blocks. Consequently, the valid data are moved to the D-blocks and the invalid pages are reclaimed. The figure 5 shows the situation before (above the arrow) and after (below the arrow) the space reclamation. There are two D-blocks: 1, 2 and two U-blocks: 8, 9, where the data from one attribute reside. Every block has 4 pages. The "shaded" pages are obsolete and the others are valid. Please note that pages 2 and 7 were updated twice: first they were written to the block 8, next to the block 9. After the reclamation, all the valid pages from the U-blocks are merged with the D-blocks so that the original order of the data is preserved. Two new D-blocks are created: 11, 12 and the previously used blocks are erased. As a result, the P-Map may be removed because S-Map is sufficient for the traverse through the data.

Let $R(a)$ and $P(s)$ denote the set of D-blocks for reclamation of the attribute a and the pages in the P-Map connected with the D-block s, respectively.

The algorithm below packs the data stored in the flash memory pages and writes them to the new D-blocks. To do this, the algorithm traverses through the pages in the D-blocks, which are assigned to be reclaimed, and the pages in the P-Maps, which are connected with them. The data are loaded to the main memory and, when their size is optimal with reference to the size of the flash memory block, they are stored to the new

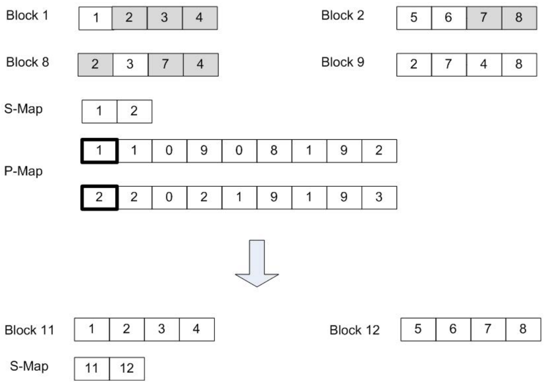


Fig. 5. Space Reclamation

D-block. It may occur the situation when for the block s the total size of the pages in $P(s)$ is less than the size of the block. In this case, the data from the next block can be appended.

Algorithm 1. PackData(INPUT $Attribute : a$)

```
Require: $PageNmb$ - the number of the page in one block
Require: $PageSize$ - the size of the page
  $data$:=$null$
  while $s \in R(a)$ do
    for $p \in P(s)$ do
      if $Size(data) + Size(p) >= PageNmb * PageSize$ then
        Write($NewBlock$):=$data$
      else
        $data$ := Append(p)
      end if
    end for
    $s$:=GetNextBlock(R(a))
  end while
  Update($P - Map$)
  Update($S - Map$)
```

3 Evaluation

In this section, we do some calculations to show the effectiveness of the column oriented model. We compare the number of pages used to answer the query for the column oriented model with the row oriented model. To do our calculation objective, we used the similar translation layer called the Row Flash Translation Layer (RFTL). It holds a collection of blocks and pages for data. It is worth noting that the contiguous blocks and pages maintain the data from the whole row, not just the data from columns as it is in case of the CFTL.

Let:
x - the number of columns in the table
y - the number of records in the table
r - the page size
$z(i)$ where $i \in \{1, 2, .., x\}$ - the size of the column
E_r - the number of rows in one memory page in the row oriented model
$E_c(z)$ - the number of z size elements in one memory page in the column oriented model
S_r - the total number of pages in case of the row oriented model
S_c - the total number of pages in case of the column oriented model

To simplify, we assume that the row may not be stored in more than one memory page. After some calculations, we get the following formulae:

$$E_r = \left\lfloor \frac{r}{\sum_{i=1}^{x} z(i)} \right\rfloor \tag{1}$$

$$S_r \leq \left\lceil \frac{y}{E_r} \right\rceil \tag{2}$$

$$E_c(z) = \left\lfloor \frac{r}{z} \right\rfloor \tag{3}$$

$$S_c = \sum_{i=1}^{x} \left\lceil \frac{y}{E_c(z(i))} \right\rceil \tag{4}$$

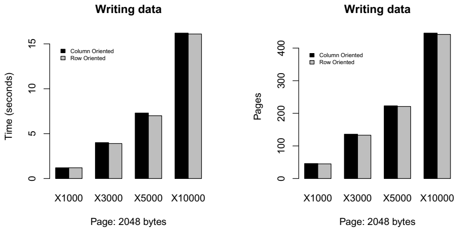
To sum up, we claim that when all the attributes of the table are queried, we have: $S_c = S_r$. The situation changes, if the query does not take into consideration all the attributes. In the row oriented model, all the attributes must be fetched whilst in the column oriented model only the relevant ones. So, the following formula holds: $S_c \leq S_r$.

4 Experiments

In this section, we shall discuss some simulations, which confirm the effectiveness of the proposed methods. We compared the time of reading and writing and the number of memory pages used for both data models. The experiments were conducted on the $1GB$ flash memory for two page sizes: 512 and 2048 bytes. We used the table $TestTab$ containing the following attributes: $A(10), B(30), C(15), D(10), E(25)$ (in brackets we provide the size of the column).

4.1 Data Writing

In this subsection, we inserted 1000, 3000, 5000 and 10000 records to the database. We repeated the experiments 100 times. In the figure 6, we present the average values resulted from the experiments. We see that time and the number of used pages are similar in both models. The inserted rows are stored in the memory buffer and then written in batches to the flash memory. Therefore, both methods need similar time and number of pages to perform bulk inserts. The situation will be different, if a single row is written to the flash memory. In that case, the row oriented model will need less time and less pages than the column oriented model to perform this operation.



(a) (b)

Fig. 6. Time and the number of pages during writing

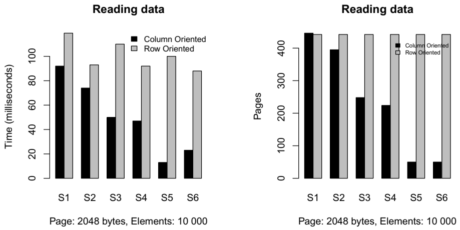
4.2 Data Reading

The aim of the following experiments is to obtain the response time and the number of read pages after processing SQL-like queries for both data models. Our experiments are conducted on the database containing 1000, 3000, 5000, 10000 records. The figure 7 shows the experimental results only for the database with 10000 records.

In our experiments we use the following SQL queries:

```
S1: SELECT A, B, C, D, E FROM TestTab
S2: SELECT A, B, C, D, E FROM TestTab WHERE C>1000 AND
C<10000
S3: SELECT A, C, E FROM TestTab
S4: SELECT A, C, E FROM TestTab WHERE C>1000 AND C<10000
S5: SELECT A FROM TestTab
S6: SELECT A FROM TestTab WHERE A>1000 AND A<10000
```

From these experiments, we provide the following conclusions. The query $S1$ selects all data from the whole table. In this case, the measured values are similar in both methods. The query $S2$ takes the whole rows, but it is limited by the range condition. We see that time and the number of pages are similar in both cases. The biggest difference can be seen in case of the queries $S5$, $S6$. It results from the fact that these queries need only data from particular columns. Therefore, the column oriented database accesses only these pages where the data for these columns reside. The column oriented model may be surely useful for the analytical databases, where such query types are used very often.



(a) (b)

Fig. 7. Time and the number of pages during reading

5 Conclusions and Future Work

In this paper, we propose the Column Flash Translation Layer as a very useful tool for the column-wise data storage. This structure enables to manage the flash memory efficiently and overcomes its limitations as far as the column oriented database are concerned. Moreover, it facilitates to fetch the data without accessing the irrelevant blocks and pages of the flash memory. We show how to efficiently store the data in the flash memory environment. Besides, we describe the data modification methods and present the precise algorithm for the space reclamation. We conduct several experiments which compare the column oriented model with the row oriented model. Their results confirm the usefulness of our approach. This paper may be treated as a fundamental description of the storage method for the column oriented database system in flash memory architecture. In this context many other aspects might be also investigated. That is why, in the future we plan to build the prototype of the fully functional DBMS considering: the query execution, the concurrency control, the join indexes, the data compression and other aspects.

Acknowledgment. This paper is partially supported by grant S20027/I-18.

References

1. Cho, H., Shin, D., Eom, Y.I.: Kast: K-associative sector translation for nand flash memory in real-time systems. In: DATE 2009, pp. 507–512 (2009)
2. Copeland, G.P., Khoshafian, S.N.: A decomposition storage model (1985)
3. Boncz, P.A., Zukowski, M., Nes, N.: Monetdb/x100: Hyper-pipelining query execution. In: CIDR, pp. 225–237 (2005)
4. Stonebraker, M., Abadi, D.J., Batkin, A., Chen, X., Cherniack, M., Ferreira, M., Lau, E., Lin, A., Madden, S., O'Neil, E., O'Neil, P., Rasin, A., Tran, N., Zdonik, S.: C-store: a column-oriented dbms. In: Proceedings of the 31st International Conference on Very Large Databases, VLDB 2005, pp. 553–564. VLDB Endowment (2005)

Parallelizing the Improved Algorithm for Frequent Patterns Mining Problem

Thanh-Trung Nguyen, Bach-Hien Nguyen, and Phi-Khu Nguyen

Department of Computer Science, University of Information Technology,
Vietnam National University HCM City, Vietnam
nguyen_thanh_trung_key@yahoo.com.vn, 08520541@aep.uit.edu.vn,
khunp@uit.edu.vn

Abstract. Mining frequent pattern has been studied for a long time. There were many algorithms introduced and proved their efficiency. But most of them have to rebuild the frequent patterns every time when there are some changes (insert, update or delete) in dataset. Accumulated Frequent Pattern has been introduced recently. It updates existing frequent patterns when there are any changes. But the time complexity is so high. This paper introduces two ways to parallelize the Accumulated Frequent Pattern algorithm and reduce the time complexity.

Keywords: Accumulated Frequent Pattern, Horizontal Parallelization, Vertical Parallelization, representative set, data mining.

1 Introduction

Frequent Patterns are set of items which occur in dataset. Finding frequent patterns is important in making decisions, generating rules and predictions. Frequent pattern problem is studied recently. Methods of finding frequent patterns fall into 3 types: Candidate Generation Methods, Without Candidate Generation Methods and Parallel Methods.

Candidate Generation Methods are almost based on Apriori algorithm [9] [13] such as pattern based algorithms and Incremental Apriori based algorithms [12], HFPA with few adaptation in Apriori for improving the interestingness of the rules produced and for applicability for web usage mining [7]. While Candidate Generation Methods focus on Apriori algorithm and improve it, Without Candidate Generation Methods focus on creating frequent pattern tree such as FP-Tree [5][11], FP-Growth [4], FP-FOREST [6], FP-Updating tree [14], DP-tree [15]… Parallelization is a new recent approach which applies advantages of machines' computation [3]. Parallelizing Apriori [8], FP-Tree[4][11] are researched recently. Grid environments are studied in order to overcome the computational limits of the original serial frequent pattern algorithm [2].

In a recent study, an Accumulated Frequent Pattern algorithm is introduced [1]. This algorithm collects and stores all possible frequent patterns in data set. The advantages of this method are easy to implement and user does not need to regenerate the frequent pattern list similar to Apriori or other frequent pattern algorithms when data are updated. But the complexity of this algorithm is too high (exponential time

A. Selamat et al. (Eds.): ACIIDS 2013, Part I, LNAI 7802, pp. 156–165, 2013.

$n2^{2m}$). This issue can make the machine run too long. Time consuming is a big problem and parallelization is one of a good way to reduce the running time of algorithm. Parallelization was applied to find frequent pattern from huge database in the past [10]. The large amount of frequent patterns is a reason makes us to apply parallelization method in Accumulated Frequent Pattern algorithm.

One of the big issues when developing an algorithm in paralleled systems is the complexity of algorithms. Some algorithms cannot divide into small part to run simultaneously in separate sites or machines. Fortunately, the accumulated frequent pattern algorithm can be expanded for parallel systems easily [1]. This paper introduces two ways to parallelize the algorithm: Horizontal Parallelization and Vertical Parallelization. The Parallelization Methods share the resource of machines and reduce the running time. It's one of the efficient ways to increase the speed of executing high complexity algorithms.

2 Accumulated Frequent Pattern Algorithm

Definition 1 (*bit-chain*): $(a_1a_2 \ldots a_m)$ (for $a_i \in \{0, 1\}$) is a m-tuple bit-chain. Zero chain is a bit-chain with each bit equals 0.

Definition 2 (*cover operation* $\hookrightarrow$): Given 2 bit-chains with the same length: $a = (a_1a_2 \ldots a_m)$, $b = (b_1b_2 \ldots b_m)$. a is said to cover b or b is covered by a – denoted $a \hookrightarrow b$ – if $pos(b) \subseteq pos(a)$ where $pos(s) = \{i \mid s_i = 1\}$

Definition 3 (*maximal pattern*)

- u is a bit-chain, k is a non-negative integer, $[u; k]$ will be a *pattern*.
- S is the set of m-tuple bit-chains (chains with the length of m), u is a m-tuple bit-chain. If there are k chains in S cover u, we say: u is a *form* of S with frequency k; and $[u; k]$ is a pattern of S – denoted $[u; k]_{\rightarrow S}$. For instance: $S = \{(1110), (0111), (0110), (0010), (0101)\}$ and $u = (0110)$. We say u is a form with frequency 2 in S, so $[(0110); 2]_{\rightarrow S}$.
- A pattern $[u; k]$ of S is called *maximal pattern* – denoted $[u; k]_{\max\rightarrow S}$ – if and only if it doesn't exist k' such that $[u; k']_{\max\rightarrow S}$ and $k' > k$. With the above instance, $[(0110); 3]_{\max\rightarrow S}$.

Definition 4 (*representative set*): P is *representative set* of S when $P = \{[u; p]_{\max\rightarrow S} \mid \nexists [v; q]_{\max\rightarrow S} : (v \hookrightarrow u \text{ and } q > p)\}$

Definition 5 (*representative pattern*): Each element in representative set P is called a *representative pattern* of S.

Let S be a set of n m-tuple bit-chains with representative set P. The *NewRepresentative* algorithm will rebuild the representative set when a new chain is added to S. [1]

```
ALGORITHM NewRepresentative (P, z)
// Finding new representative set for S when
                    one bit-chain is added to S.
// Input: P is a representative set of S,
          z is a bit-chain added to S.
// Output: The new representative set P of $S \cup \{z\}$.
M = $\varnothing$ // M: set of new elements of P
flag1 = 0
flag2 = 0
for each $x \in P$ do
  $q = x \circ [z; 1]$
  if $q \neq 0$ // q is not a bit-chain with all bits 0
    if $x \subseteq q$ then $P = P \setminus \{x\}$
    if $[z; 1] \subseteq q$ then flag1 = 1
    for each $y \in M$ do
      if $y \subseteq q$ then
        $M = M \setminus \{y\}$
        break for
      endif
      if $q \subseteq y$ then
        flag2 = 1
        break for
      endif
    endfor
  else
    flag2 = 1
  endif
  if flag2 = 0 then $M = M \cup \{q\}$
  flag2 = 0
endfor
if flag1 = 0 then $P = P \cup \{[z; 1]\}$
$P = P \cup M$
return P
```

3 The Principles of Parallel Computing

Parallel computing is a form of computation in which many calculations are carried out simultaneously operating on the principle that large problems can often be divided into smaller ones, which are then solved concurrently.

The parallel system of accumulated frequent pattern algorithm has this structure:

At first, the fragmentation will be implemented. The whole data will be divided into small and equal fragments. In Horizontal Parallelization, the tuples (records) in data will be divided while the attributes are the information which will be divided in Vertical Parallelization. After the data is fragmented properly, all fragments must be allocated in various sites of network. A master site has responsibility for fragmenting data, allocating fragments to sites and merging the results from site into the final result. Sites will run the Accumulated Frequent Pattern algorithm [1] with the fragments which are assigned to them simultaneously and finally, send the results into the master site for merging.

4 Horizontal Parallelization

Consider a set of *n* invoices *S*. If there are *k* machine located in k separate sites. Horizontal Parallelization (HP) will divide the set *S* into *k* equal fragments and allocate those fragments to *k* sites. The Accumulated Frequent Pattern algorithm is applied on *n*/*k* invoices in each site. After running algorithm, every site has its own representative set. All representative sets are sent back to master site for merging.

Merging representative sets from sites in mater site is similar to find representative set when adding new invoices into dataset.

```
ALGORITHM HorizontalMerge(PM, PS)
//Input: PM: representative set of master site.
         PS: the set of representative sets from other
                                                  sites.
//Output: The new representative set of horizontal
                                        parallelization.
for each P ∈ PS do
  for each z ∈ P do
    PM = NewRepresentative(PM, z);
  end for;
end for;
return PM;
```

Theorem 1: Horizontal Parallelization method returns the representative set.

Proof: We prove by induction on *k*.

If $k = 1$, then the HP method is the *NewPresentative* method, hence returns the representative set.

If $k = 2$, then let the two sites be S_1 and S_2, and let (u_1, p_1), … , (u_m, p_m) be the representative set for S_1, and let (v_1, q_1), … , (v_n, q_n) be the representative set for S_2. We need to show that the HP algorithm will give us the representative for the union of S_1 and S_2. Let (w, r) be a representative element for the union of S_1 and S_2. We denote by (w_1, r_1) the restriction of (w, r) to S_1, that is $w_1 = w$ and r_1 = the number of when *w* appears in S_1. Similarly we define (w_2, r_2). Then by definition we must have $r_1 + r_2 = r$. Also, there must be one of the representative elements in S_1, called (u_1, p_1), so that *w* is contained in u_1 and $p_1 = r_1$. In fact, by definition of representative elements, there must be such a (u_1, p_1) for which *w* is contained in u_1 and $p_1 >= r_1$. We also have a (v_1, q_1) in S_2 with *w* is contained in v_1 and $q_1 >= r_2$. Now we must have $p_1 = r_1$ and $q_1 = r_2$ because otherwise *w* will appear $p_1 + q_1 > r_1 + r_2 = r$ times in the union of S_1 and S_2. Now when we apply the HP algorithm then we will see at least one element (w', r) where w' = the intersection of u_1 and v_1, and in particular w' must contain *w*. Then w' must be in fact *w*, otherwise, we have an element (w', r) with w' contains *w* but is strictly larger than *w*, and hence (w, r) cannot be a representative element. Then we see that (w, r) is produced when using the HP algorithm as wanted.

Assume that we proved the correctness of the HP algorithm for *k* sites. We now prove the correctness of the HP algorithm for $k + 1$ sites. We denote these sites by S_1, S_2, … , S_k, S_{k+1}. By the induction assumption, we can find the representative set for *k* sites S_1, S_2, … , S_k using the HP algorithm. Denote by S = the union of S_1, … , S_k. Now apply the case of 2 sites which we proved above to the two sites S and S_{k+1}, we have that the HP algorithm produces the representative set for the union of S and S_{k+1}, that is we have the representative set for the union of the $k + 1$ sites S_1, … , S_k, S_{k+1}. Therefore we completed the proof of Theorem 1.

5 Vertical Parallelization

Vertical Parallelization is more complex than Horizontal Parallelization. While Horizontal Parallelization focuses on tuples (invoices), Vertical Parallelization focuses on attributes (goods). Vertical Parallelization (VP) divides the dataset into fragments based on attributes. Each fragment contains a subset of attributes. Fragments are allocated into separate sites. In each sites, they run *NewRepresentative* algorithm to find out representative sets. The representative sets will be sent back to the master site and will be merged to find the final representative set.

Vertical merging is based on merged-bit operation and vertical merging operation.

Definition 6 (*merged-bit operation* $\uplus$): *merged-bit operation* is a dyadic operation in bit-chain space $(a_1a_2 \dots a_n) \uplus (b_1b_2 \dots b_m) = (a_1a_2 \dots a_nb_1b_2 \dots b_m)$.

Definition 7 (*vertical merging operation* $\nleftrightarrow$): *Vertical merging operation* is a dyadic operation between two representative patterns from two different vertical fragments:

$[(a_1a_2 \dots a_n), p, \{o_1, o_2, \dots , o_p\}] \nleftrightarrow [(b_1b_2 \dots b_m), q, \{o_1, o_2, \dots, o_q\}] = [vmChain, vmFreq, vmObject]$;

$vmChain = (a_1a_2 \dots a_n) \uplus (b_1b_2 \dots b_m) = (a_1a_2 \dots a_nb_1b_2 \dots b_m)$

$vmObject = \{o_1, o_2 , \dots , o_p\} \cap \{o_1, o_2 , \dots , o_q\} = \{o_1, o_2, \dots , o_k\}$

$vmFreq = k$

At the master site, representative sets of other sites will be merged into representative set of master site. The below algorithm is run to find out the final representative set.

```
ALGORITHM VerticalMerge(PM, PS)
//Input: PM: representative set of master site.
         PS: the set of representative sets from other
                                                   sites.
//Output: The new representative set of vertical
                                    parallelization.
for each P ∈ PS do
  for each m ∈ PM do
    flag = 0;
    M = ∅ // M: set of used elements in P
    for each z ∈ P do
      q = m ↮ z;
      if frequency of q ≠ 0 then
        flag = 1; //mark m as used element
        M = M ∪ {q}; //mark q as used element
```

```
          PM = PM ∪ {q};
        end if;
      end for;
      if flag = 1 then
        PM = PM ∪ {m}
      end if;
      PM = PM ∪ (P \ M);
    end for;
  end for;
  return PM;
```

Theorem 2: Vertical Parallelization method returns the representative set.

Proof: We prove by induction on k.

If k = 1, this is the New Representative algorithm, hence gives the representative set.

If k = 2, we let S_1 and S_2 be the two sites, and let S be the union. Let $R = (\{a_1, \dots, a_n\}, k, \{o_1, \dots, o_k\}) = (A, k, O)$ be a representative element for S. Let R_1 be the restriction of R to S_1, that is $R_1 = (A_1, p_1, O_1)$, where A_1 is the intersection of $\{a_1, \dots, a_n\}$ with the set of attributes in S_1, O_1 is the set of invoices in R_1 containing all attributes in A_1. We define similarly $R_2 = (A_2, p_2, O_2)$ the restriction of R to S_2. Note that if A_1 is empty then p_1 is zero, otherwise p_1 must be positive. Similarly, if A_2 is empty then p_2 is zero, otherwise p_2 must be positive. We use the convention that if A_1 is empty then $O_1 = O$, and if A_2 is empty then $O_2 = O$. Remark that at least one of A_1 or A_2 is non-empty. Now by definition, there must be a representative element $Q_1 = (A_1', p_1', O_1')$ in S_1 with A_1' contains A_1, O_1' contains O_1, and $p_1' >= p_1$. Similarly, we have a representative element $Q_2 = (A_2', p_2', O_2')$ in S_2 with A_2' contains A_2, O_2' contains O_2, and $p_2' >= p_2$. When we do the vertical merging operation of Q_1 and Q_2, then we must obtain A. Otherwise, we obtain an element $R' = (A', k', O')$ where A' contains A, O' contains O, and $k' >= k$, and at least one of these is strict. This is a contradiction to the assumption that R is a representative element of S.

Now that we proved the correctness of the algorithm for k = 1 and k = 2, we can follow the prove in the same way as in the end of the proof of Theorem 1 to finish the proof of Theorem 2.

Example

Give the set S of invoices $\{o_1, o_2, o_3, o_4, o_5, o_6, o_7\}$ and the set I of goods $\{i_1, i_2, i_3\}$.

	i_1	i_2	i_3
o_1	0	0	1
o_2	0	1	0
o_3	0	1	1
o_4	1	0	0
o_5	1	0	1
o_6	1	1	0
o_7	1	1	1

Fig. 1. Bit-chains of S

Divide the attributes (goods) into two segments: $\{i_1\}$ and $\{i_2, i_3\}$.

	i_1
o_1	0
o_2	0
o_3	0
o_4	1
o_5	1
o_6	1
o_7	1

Fig. 2. The first segment

	i_2	i_3
o_1	0	1
o_2	1	0
o_3	1	1
o_4	0	0
o_5	0	1
o_6	1	0
o_7	1	1

Fig. 3. The second segment

Allocate two segments in two separate sites and run *NewRepresentative* algorithm on those fragments simultaneously. It is easy to get the representative sets from two sites:

Sites has $\{i_1\}$ fragment: $P_{\{i_1\}}=\{[(1); 4; \{o_4, o_5, o_6, o_7\}]\}$

Site has $\{i_2, i_3\}$ fragment: $P_{\{i_2,i_3\}}=\{[(01); 4; \{o_1, o_3, o_5, o_7\}], [(10); 4; \{o_2, o_3, o_6, o_7\}], [(11); 2; \{o_3, o_7\}]\}$

Merge both of them:

$[(1); 4; \{o_4, o_5, o_6, o_7\}]$ ⇹ $[(01); 4; \{o_1, o_3, o_5, o_7\}] = [(101); 2; \{o_5, o_7\}]$
$[(1); 4; \{o_4, o_5, o_6, o_7\}]$ ⇹ $[(10); 4; \{o_2, o_3, o_6, o_7\}] = [(110); 2; \{o_6, o_7\}]$
$[(1); 4; \{o_4, o_5, o_6, o_7\}]$ ⇹ $[(11); 2; \{o_3, o_7\}] = [(111); 1; \{o_7\}]$

The full representative set P is:

$P = \{[(100); 4; \{o_4, o_5, o_6, o_7\}],$
$[(001); 4; \{o_1, o_3, o_5, o_7\}],$
$[(010); 4; \{o_2, o_3, o_6, o_7\}],$
$[(011); 2; \{o_3, o_7\}],$
$[(101); 2; \{o_5, o_7\}],$
$[(110); 2; \{o_6, o_7\}],$
$[(111); 1; \{o_7\}]\}$

6 Experimentation

The experiments of Vertical Parallelization Method are conducted on machines which has similar configuration: Intel(R) Core(TM) i3-2100 CPU @ 3.10GHz (4 CPUs), ~3.1GHz and 4096MB main memory installed. The operating system is Windows 7 Ultimate 64-bit (6.1, Build 7601) Service Pack 1. Programming language is C#.NET.

The proposed methods are tested on T10I4D100K dataset taken from *http://fimi.ua.ac.be/data/* website [1]. First 10,000 invoices of T10I4D100K are run and compared with the result when running without parallelization.

Table 1. The result when running in single machine

Dataset	No. of invoices	The maximum No. of goods which customer can purchase	The maximum No. of goods in the dataset	Running time	No. of frequent patterns
T10I4D100K	10,000	26	1,000	702.13816s	139,491

Vertical Parallelization model is applied in 17 machines with the same configuration. Each machine is located in a separate site. 1,000 goods is digitized into 1,000-tuple bit-chain. The master site will divide the bit-chain into 17 fragments (16 60-tuple bit-chains and a 40-tuple bit-chain).

Table 2. The result of running algorithm in sites

Fragments	Length of bit-chain	Running Time	Number of frequent patterns
1	60	18.0870345s	899
2	60	10.5576038s	535
3	60	14.9048526s	684
4	60	12.2947032s	548
5	60	8.0554607s	432
6	60	10.3075896s	560
7	60	17.7480151s	656
8	60	10.8686217s	526
9	60	21.3392205s	856
10	60	13.3007607s	682
11	60	9.6135499s	617
12	60	16.1179219s	736
13	60	15.4338827s	587
14	60	21.4922293s	928
15	60	18.2790455s	834
16	60	17.011973s	701
17	40	4.4142525s	223

After running algorithm in sites, the representative sets are sent back to master site for merging. The merging time is 27.5275745s and the number of final frequent patterns is 139,491. So, the total time of Parallelization is:

max(Running Times) + Merging Time = 21.3392205s + 27.5275745s = 48.866795s

It is easy to see the efficiency and accuracy of Vertical Parallelization method.

7 Conclusion

Mining frequent pattern has been researched much recently. Accumulated Frequent Pattern algorithm was introduced and proved its advantages [1]. But the complexity is a big problem of this algorithm. The paper presented two methods to reduce the complexity of Accumulated Frequent Pattern algorithm: Vertical Parallelization and Horizontal Parallelization methods. The theory base and experimental results prove the efficiency of these methods.

References

1. Nguyen, T.-T., Nguyen, V.-L.H., Nguyen, P.-K.: Accumulated Frequent Pattern – ICTMF2012: The Third International Conference on Theoretical and Mathematical Foundations of Computer Science, Bali, Indonesia, December 1-2 (2012)
2. Appice, A., Ceci, M., Turi, A., Malerba, D.: A parallel, distributed algorithm for relational frequent pattern discovery from very large datasets. Intelligent Data Analysis 15, 69–88 (2011)
3. Niimi, A., Yamaguchi, T., Konishi, O.: Parallel Computing Method of Extraction of Frequent Occurrence Pattern of Sea Surface Temperature from Satellite Data. In: The Fifteenth International Symposium on Artufical Life Robotics (AROB 15th 2010), B-Con Plaza, Beppu, Oita, Japan, February 4-6 (2010)
4. Xiaoyun, C., Yanshan, H., Pengfei, C., Shengfa, M., Weiguo, S., Min, Y.: HPFP-Miner: A Novel Parallel Frequent Itemset Mining Algorithm. In: Fifth International Conference on Natural Computation, ICNC 2009, August 14-16 (2009)
5. Yen, S.-J., Wang, C.-K., Ouyang, L.-Y.: A Search Space Reduced Algorithm for Mining Frequent Patterns. Journal of Information Science and Engineering 28, 177–191 (2012)
6. Li, H., Zhang, N., Chen, Z.: A Simple but Effective Maximal Frequent Itemset Mining Algorithm over Streams. Journal of Software 7(1) (January 2012)
7. Sudhamathy, G., Venkateswaran, C.J.: An Efficient Hierarchical Frequent Pattern Analysis Approach for Web Usage Mining. International Journal of Computer Applications 43(15), 0975–8887 (2012)
8. Verma, G., Nanda, V.: Frequent Item set Generation by Parallel Preprocessing on Generalized Dataset. International Journal of Scientific & Engineering Research 3(4) (April 2012)
9. Nancy, P., Ramani, R.G.: Frequent Pattern Mining in Social Network Data (Facebook Application Data). European Journal of Scientific Research 79(4), 531–540 (2012) ISSN 1450-216X

10. Gupta, R., Satsangi, C.S.: An Efficient Range Partitioning Method for Finding Frequent Patterns from Huge Database. International Journal of Advanced Computer Research 2(2) (June 2012) (ISSN (print): 2249-7277 ISSN (online): 2277-7970)
11. Bhadane, C., Shah, K., Vispute, P.: An Efficient Parallel Approach for Frequent Itemset Mining of Incremental Data. International Journal of Scientific & Engineering Research 3(2) (February 2012)
12. Duneja, E., Sachan, A.K.: A Proficient Approach of Incremental Algorithm for Frequent Pattern Mining. International Journal of Computer Applications (0975 – 888) 48(20) (June 2012)
13. Miura, T., Okada, Y.: Extraction of Frequent Association Patterns Co-occurring across Multi-sequence Data. In: Proceedings of the International MultiConference of Engineers and Computer Scientists, IMECS 2012, Hong Kong, March14-16 (2012)
14. Baskar, M.S.S.X., Dhanaseelan, F.R., Christopher, C.S.: FPU-Tree Frequent Pattern Updating Tree. International Journal of Advanced and Innovative Research 1(1) (June 2012)
15. Hafija, S., Murthy, J.V.R., Anuradha, Y., Sekhar, M.C.: Mining Frequent Patterns from Data streams using Dynamic DP-tree. International Journal of Computer Applications (0975 – 8887) 52(19) (August 2012)

Dimensionality Reduction in Data Summarization Approach to Learning Relational Data

Chung Seng Kheau, Rayner Alfred, and Lau Hui Keng

School of Engineering and Information Technology, Universiti Malaysia Sabah, Jalan UMS,
88400, Kota Kinabalu, Sabah, Malaysia
{kheau,ralfred,hklau}@ums.edu.my

Abstract. Due to the growing amount of digital data stored in relational databases, more new approaches are required to learn relational data. The DARA algorithm is designed to summarize data and it is one of the approaches introduced in relational data mining in order to handle data with one-to-many relations. The DARA algorithm transforms data stored in relational databases into a vector space representation by applying the information retrieval theory. Based on the experimental results, the DARA algorithm is proven to be very effective in learning relational data. However, DARA suffers a major drawback when the cardinalities of attributes are very high because the size of the vector space representation depends on the number of unique values that exist for all attributes in the dataset. This paper investigates the effects of discretizing the magnitude of terms computed and applying a feature selection process that reduces the cardinalities of attributes of the relational datasets on the predictive accuracy of the overall classification task. This involves the task of finding the best set of relevant features used to summarize the data, in which the feature selection processed is performed based on the magnitude of terms computed earlier. Based on the results obtained, it shows that the predictive accuracy of the classification task can be improved by improving the quality of the summarized data. The quality of the summarized data can be enhanced by appropriately discretizing the magnitude of terms computed earlier and also appropriately selecting only a certain percentage of the attributes.

Keywords: Relational Data Mining, Data Summarization, Clustering, Dimensionality Reduction, Discretization Numbers, Feature Selection.

1 Introduction

Nowadays, most scientific data are stored in relational databases. A relational database describes a collection of data items that are stored in a set of relations, also known as tables, in which the data can be accessed and manipulated easily. A relation is a data structure which consists of columns that represent attributes and rows that represent tuples in a form of table. A tuple normally represents an object with one or more attributes information. An attribute is represented by a specific information of the tuple. However, in a relational database, a single tuple (record) in a specific table can be

A. Selamat et al. (Eds.): ACIIDS 2013, Part I, LNAI 7802, pp. 166–175, 2013.

associated with multiple tuples (records) stored in another table (as shown in Fig. 1). This is known as a one-to-many relationship among records stored in multiple tables.

This paper is organized as followed. Section 2 explains some works related to relational data mining. Section 3 describes the general overview of the DARA algorithm process. Section 4 discusses the data transformation applied in DARA. Section 5 discusses the experimental setup and the results obtained in investigating the effects of discretizing the magnitude of terms computed and applying a feature selection process that reduces the cardinalities of attributes of the relational datasets on the predictive accuracy of the overall classification task. This paper is concluded with future works in Section 6.

2 Relational Data Mining Techniques

Given data stored in relational databases, traditional data mining techniques cannot be applied directly in order to learn the stored data. Therefore, several approaches have been proposed to learn relational data.

A Multi-relational Bayesian Classification Algorithm with Rough Set (RS-RBC) [15] introduces the concept of using relational graph to dynamically choose the associative table and its associated table. It uses a tuple ID propagation approach to overcome the association rule mining problem with multiple database relations. The concept of Core in Rough Set is to simplify the associative table. This algorithm is able to directly support the relational database and its running rate is much higher than ILP (Inductive Logic Programming) based Multi-Relational Classification Algorithm.

ILP [20], also known as logic-based MRDM (Multi-relational data mining), covers area of researches that falls under the intersection of machine learning and logic programming. It is characterized by using logic to represent the multi relational data. The ILP-Based Relational Classification approach searches for syntactically legal hypotheses constructed from the predicates and predict the class labels based on examples in background knowledge. There are three main categories: decision trees relational classification, distance relational classification (RIBL [21,22] and Kernel [23]) and probability classification approach (PRM [24] and SLP [14]).

Decision trees relational classification classifies the instances by sorting them into tree structure based on their feature values. Instances are classified starting from the root of a node based on their sorting feature values. Each node in the tree represents a feature in an in instance, and each branch represents a value for node to assume. Decision tree is appropriate for exploratory knowledge because its construction does not require any domain knowledge. Generally, decision tree classifiers have good accuracy.

In distance relational classification approach, the RIBL (Relational Instance-Based Learning) [21,22] learns the distance (similarity) between objects in the multi-relational database environment. First, at depth 0, the objects' properties are taken into account to calculate the distance between two objects. Next, objects that are related to the two original objects are taken into account (at depth 1). At depth 2, the objects that related to those at depth 1 are taken into account, and so on, until the user-specified depth limit is reached. RIBL uses a *k*-nearest-neighbor algorithm in conjunction with the RIBL distance measure to overcome the prediction problem. Another type of distance-based relational classification called a kernel function is designed to project data in non-linear space into high dimensional linear hyper sphere feature spaces and then classify the data based on their distances. They apply direct sum kernel and kernel which is derived from the R-Convolution kernel [23]. Probability classification approaches integrate both logical and probabilistic approaches to knowledge representation and reasoning. Probabilistic relational model (PRM) [24] is an extension of Bayesian networks that uses the relational structure for handling relational data. The learning task is to extract the good dependency structure from the training database. It specifies and evaluates the candidate hypotheses and carries out refinement search in the hypothesis space for a high score structure. Stochastic Logic Programs (SLPs) [14] is the integration of Stochastic Grammars with logic. It consists of a set labeled clauses *p:C*, in which *p* is a probability label describing the information that corresponding to relational pattern, and *C* is a logic clause for extended dependent relationship among the data. The clause set covers each specific example and the probabilities record the dependence relationships, when learning the data.

In Karunaratne *et al.* [13], graph propositionalization methods can be used as a preprocessing method to transform structured data into fixed-length feature vector for standard machine learning algorithms. Three propositionalization methods, that include MOSS, SUBDUE and SMFI, have been proposed in conjunction with three standard machine learning algorithms: random forests, support vector machines and nearest neighbor classifiers. Results have shown a significant impact on the result accuracy based on 21 datasets in the domain of medicinal chemistry.

Aforementioned, the increase in the amount of digital data stored in relational databases also increases the complexity of the structure in multiple tables. This has an effect on the efficiency and the accuracy of the existing classification algorithms. Therefore, Li *et al.* [16] proposed to classify the multi-relational based on the contribution of tables. Each table in the database is computed with a contribution score and sorted according to their contribution score. Then, the contribution tables with a higher score will be chosen while those of lower score will be excluded. However, adopting a static sorting algorithm only improves the efficiency of the classification because it only involves some of the tables instead of all tables. Pan *et al.* [17] proposed the reduction of attribute twice in multi-relational classification. The two pruning strategy is designed to get rid of attribute based on the foil gain.

Liau *et al.* [18] proposed an approach to select only effective features and relations for improving the efficiency of the multi-relational classification task. In this approach, a symmetrical uncertainty is applied to measure the correlation between attributes in a table or cross table, and also between a table and the class attribute. Finally only relevant attributes and tables are selected from the database for classification process.

In a multiple view strategy [19], relational objects are classified with conventional data mining methods without having to flatten the multiple relations into a universal one. Multiple view learners are employed to capture separately the essential information that embedded in the individual relation.

3 Dynamic Aggregation for Relational Attributes (DARA)

DARA is an algorithm that is particularly designed to summarize data stored in relational databases that consist of data with one-to-many relations [8, 10]. In a relational database, the *target* table (labeled data) is normally linked to other *non-target* tables with one-to-many relationships. A *target* table is said to have a one-to-many relationship with the non-target table if each record stored in the target table can be associated to one or more records stored in the non-target table (see Fig. 1). *TT1* is a *target* table. *NT1* is a *non-target* table. In this scenario, table *TT1* has a one-to-many relationship with table *NT1*, through the association of *ID1* and *ID2*. In DARA algorithm, the entire contents of the *non-target* tables are summarized according to the individual record stored in the *target* table. There are three stages involved in the summarization process which include data pre-processing, data transformation and data summarization. In the data pre-processing stage, data dimensionality can be reduced by applying some of the well know techniques in data reduction. These methods include discretisation of continuous values into categorical values, feature selection and construction [9, 10, 11]. All the processes involved in the data pre-processing stage are performed before the relational data representation is transformed into a vector space representation.

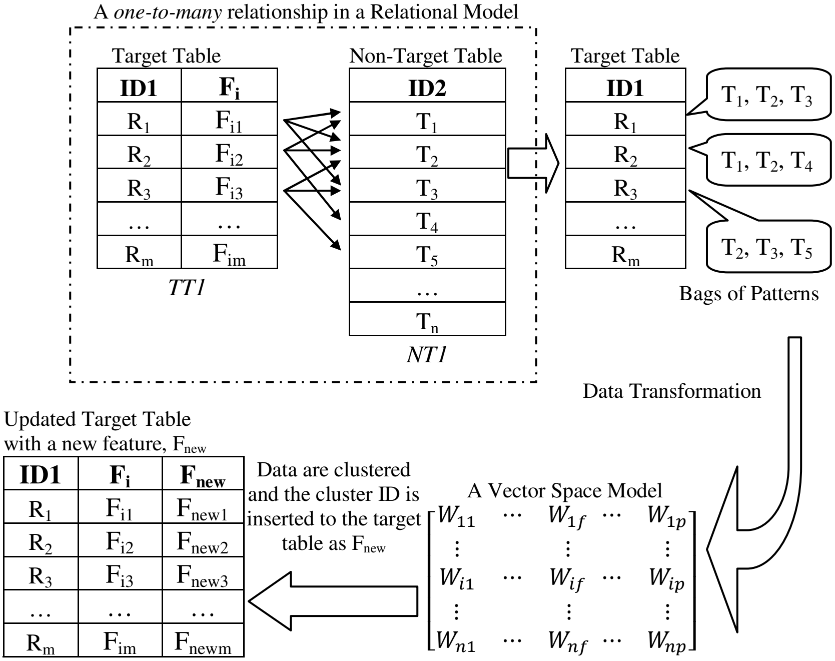


Fig. 1. Data transformation process and the creation of a new feature in the target table

The next stage is the data transformation stage that includes two main tasks; Data model conversion and computation of component magnitude for each feature extracted. In this stage, a relational data representation is transformed into a vector space model representation in which a record stored in the target table, in a relational database, that has a *one-to-many* relationship with records stored in another non-target table can be represented as a record-pattern matrix or a vector space representation (see Fig. 1). In this record-pattern matrix, a row represents a single record extracted from the non-target table in a relational database and the column represents a pattern that exists for that particular record. Once the data transformation has been performed, the data summarization process can be performed in which records stored in the non-target table are clustered and each record will be given a label to indicate the group that the records belong to. The summarized data can further be appended to the target table as an additional column or as a new feature. The updated target table can then be evaluated by using any traditional data mining techniques (e.g., C4.5 or Naïve Bayes) [1]. Based on the empirical results obtained, it is shown that DARA algorithm manages to improve the predictive accuracies of the C4.5 classifier compared to other relational data mining methods [9, 11]. However, DARA suffers a major drawback in which the vector space dimensionality will grow larger when the number of distinct values for each column in the relational database increases. This is due to the fact that during the data transformation process, each unique value found in the relational database contributes a new column in the vector space model. When the dimensionality of the vector space is large, the descriptive accuracy obtained from the data summarization process will not be accurate. Thus, this paper proposes an optimization method that will enhance the data summarization process when the vector space is large.

4 Feature Transformations in DARA

Feature transformations, that include feature construction, feature selection and feature extraction, have been used wisely in the area of pattern recognition, machine learning, statistics and data mining communities [2, 3]. The purpose of feature transformations is to reduce inputs to a manageable size for data processing and analysis, by constructing, selecting and extracting only relevant features and eliminating less predictive information. This is important because the predictive accuracy of any classification tasks depends on the quality of the input data. For this reason, a dataset normally requires some pre-processing steps before any learning takes place for data analysis purposes. Feature selection for relational datasets requires more attention compared to feature selection for dataset stored in a single table. This is due to the fact that in relational datasets, a single record stored in the target table may be associated to several records stored in a non-target table.

4.1 Feature Construction in the DARA's Transformation Algorithm

In DARA algorithm [9, 11], a feature construction method has been used to construct a set of relevant features before the data transformation from a relational data model representation to a vector space model representation takes place. The feature construction method is used as the basic foundation in order to transform relational data model representation into a vector space model representation. After the data has been transformed into a vector space model representation, the data can be represented as a record-pattern matrix. In this record-pattern matrix, a row represents a single record and the list of columns represents a series of features obtained after the data transformation process is performed. From this record-pattern matrix, a feature selection method can be further proposed in order to select a set of relevant features used before the data summarization or clustering process takes place (as shown in Fig. 2).

4.2 Feature Selection for the Record-Pattern Matrix in DARA

In this paper, a further pre-processing step (a feature selection method) is proposed to select features from the vector space model representation (e.g., the record-pattern matrix). The vector space model representation is in the form of *TF-IDF* (*term frequency-inverse document frequency*) weighted frequency matrix. In information retrieval theory, *TF-IDF* is a statistical measure which expresses the importance a word [12]. In our work, we apply the *TF-IDF* as a statistical measure which expresses the importance a feature. *TF-IDF* score becomes bigger if a feature appears a lot in a record, but it lowers if the feature also appears in many other records. As a result, before the feature selection process can be performed based on the feature scoring, all numerical values are required to be discretized in order to reduce the dimensionality of the record-pattern matrix. The data values for each feature are converted to a character based number. After the discretisation process, the score of each feature, *F*, will be computed by using the information gain measure, *InfoGain(F)*, as shown in Equation 1. Given a particular feature *F*, the information gain for this feature, denoted *InfoGain(F),* represents the difference between the class entropy in data set before the usage of feature *F*, denoted *Ent(C),* and the usage of feature *F* with splitting the data set into subsets, denoted *Ent(C/F)*, as presented in Equation 1.

$$InfoGain(F) = Ent(C) - Ent(C|F) \tag{1}$$

where

$$Ent(C) = -\sum_{j=1}^{n} Pr(C_j) \cdot \log_2 Pr(C_j) \tag{2}$$

and

$$Ent(C|F) = -\sum_{i=0}^{n} Pr(F_i) \cdot (-\sum_{i=0}^{n} Pr(C|F_i) \cdot \log_2 Pr(C_j|F_i)) \tag{3}$$

After feature scoring is performed, all features will be ranked based on the values obtained from the feature scoring. Finally, the feature selection process can be performed to select a small percentage of features for the data summarization or clustering process. For instance, one may select top 10% of 200 features that exist in the vector space model representation (e.g., the record-pattern matrix). As a result, the new vector space model may consist of only 20 features for the data summarization or clustering process [6].

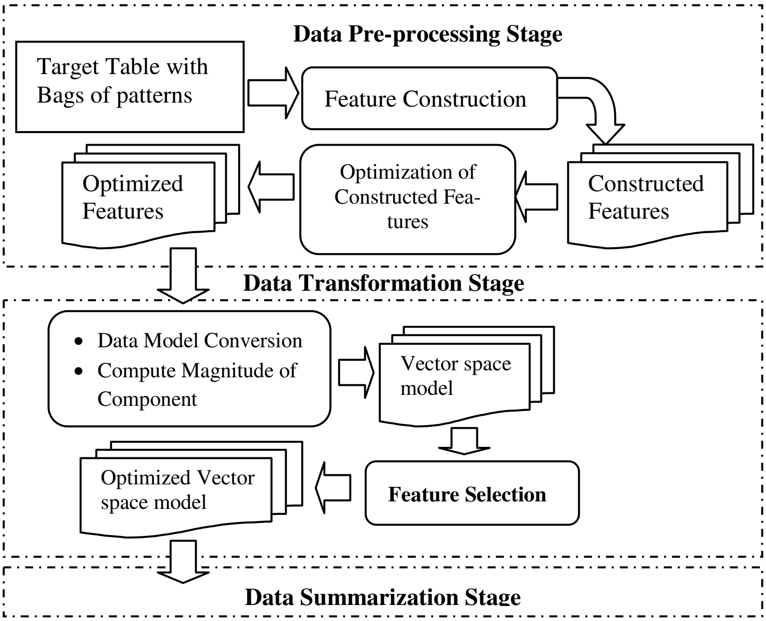

Data Pre-processing Stage
Target Table with Bags of patterns
Feature Construction
Constructed Features
Optimization of Constructed Features
Optimized Features
Data Transformation Stage
Data Model Conversion
Compute Magnitude of Component
Vector space model
Feature Selection
Optimized Vector space model
Data Summarization Stage

Fig. 2. Feature selection method is applied to the data stored as a vector space model

5 Experimental Design and Results

There are two main parts of this study. In the first part of this study, the data will be clustered or summarized based on the given number of clusters, *K*. The main task here is to find the best number of clusters in order to summarize the transformed data. The ranges of *K* will be from 3 to 13 with an increment of 2. In other words, $K = \{3, 5, 7, 9, 11, 13\}$. After finding the best number of *K*, then we proceed with the experiment by selecting the best number of features in order to summarize the data. The clustering results are then appended into the target table as a new additional feature, as shown in Fig. 1, and fed into a C4.5 classifier (J48 in *WEKA*) [7]. A 10-fold cross-validation will be used to evaluate the predictive accuracy of the classifier for the given data. It is found that the best number of clusters for classifying the Mutagenesis dataset is 7. It can be concluded that for the mutagenesis dataset, the optimum number of clusters, *K*, is 7.

In the second part of this study, the experimental design involves two main preprocessing stages. In the first stage, data that are stored in the vector space model will be discretized in order to reduce the number of unique values that exist in the data. Since, in the DARA algorithm, all data stored in a relational model are transformed into a vector space model by using the information theory concept (e.g., TF-IDF), in which all data computed and stored in the vector space model representation will have a numerical data type. For this reason, the range of values will be large. In order to reduce the cardinality of each feature in the dataset, all numerical values will be discretized according to the number of bins given (e.g., bin = 4). For instance, if the number of bins is 4, then the data for that particular feature will be divided into 4 partitions. In other words, the cardinality of that particular feature will be reduced to only four. In this experiment, the number of bins used for evaluation will be [3,5,7,9].

In the second stage, a feature selection process is performed in order to reduce the dimensionality of the dataset. The feature selection process will be conducted based on the scoring of the given feature, *F*, which is computed by using the information gain, *InfoGain*(*F*) as shown in (1). The number of features selected for the data summarization will be based on the threshold, *t*, which indicates the percentage of features used in the data summarization process. For instance, if the number of feature is *n* and the threshold for the number of features, *t*, is 0.8, then we only have (0.8 x *n*) features that will be used in the data summarization process. In this experiment, the percentage of features selected will from 20 to 100. Table 1 shows the predictive accuracy results of the C4.5 classifier on the three mutagenesis datasets (B1, B2, and B3). These predictive accuracies are based on different number of bins (ranges [3,5,7,9]) and different percentage of features involved in the data summarization process. For B1 dataset, the results indicate that the top 50% of the highest ranked features is able to get the highest predictive accuracy of 82.3% with bin = 3. For B2 dataset, the highest predictive accuracy of 82.3% is obtained when using the top 70% of the features with bin=7. Finally, for the B3 dataset, the highest predictive accuracy of 81.7% is obtained when using all features (100%) with bin = 5. In short, based on the results shown in Table 1, when we take 7 clusters as the optimum number of clusters and consider only 50% of the features given based on the feature scoring using the information gain, the predictive accuracy is improved from 81.0% 82.3 for Mutagenesis dataset B1. The same pattern can be observed in the mutagenesis dataset B2 and B3 in which the predictive accuracy can be improved from 81.6% and 81.3%, to 82.3% and 81.7% respectively.

Table 1. Performance accuracy (%) of 10-fold cross-validation of C4.5 with Feature Selection and Bin on Mutagenesis datasets (B1, B2, and B3)

Mutagenesis Datasets	**# of Bin**	**Without Feature Selection**	**S, Feature Selection (%)**							
			90	**80**	**70**	**60**	**50**	**40**	**30**	**20**
B1	3	80.4	79.2	79.5	79.5	**80.4**	**82.3**	80.2	80.2	80.6
	5	81.0	78.7	79.3	**79.8**	79.1	79.8	**81.1**	**80.8**	**80.9**
	7	80.5	78.7	**80.5**	78.2	78.8	79.9	**81.1**	**80.8**	**80.9**
	9	**81.7**	**80.7**	78.2	78.9	80.3	79.5	**81.1**	**80.8**	**80.9**
B2	3	80.6	80.4	**81.3**	80.4	**81.0**	78.2	**79.3**	80.0	79.0
	5	**81.8**	81.1	80.6	81.6	80.1	**78.7**	78.5	79.3	**79.4**
	7	81.4	80.5	81.1	**82.3**	79.9	78.3	78.3	**80.3**	78.7
	9	80.7	**81.4**	81.1	80.2	80.3	78.3	78.8	79.8	79.0
B3	3	81.1	79.3	78.0	**78.6**	79.5	**79.9**	80.2	78.8	79.6
	5	**81.7**	**81.0**	**78.7**	**78.6**	78.8	79.6	**80.6**	80.8	**80.9**
	7	80.5	80.7	78.0	78.4	**79.9**	**79.9**	80.4	80.8	**80.9**
	9	81.2	79.7	78.3	78.3	78.8	79.7	80.3	**80.9**	**80.9**

It is also observed that the data transformation process for relational datasets, that transforms this data from a relational model representation into a vector space model representation, may produce unwanted features which may affect the accuracy of the predictive task. This can be observed in Table 1, where the highest predictive accuracies obtained for Mutagenesis dataset B1 and B2 are 82.3% and 82.3 when only 50% and 70% of the features are selected respectively. The highest predictive accuracies obtained for all three datasets (B1, B2 and B3) are obtained with different percentage of features selected with different number of bins used to discretize numerical values. These results illustrate that every feature that exists in the vector space model requires different number of bins because every feature has its own range of values due to the nature of the data transformation process performed by DARA.

6 Conclusion

This paper investigates the feasibility of fine-tuning some of the parameters in DARA in learning relational datasets in order to improve the predictive accuracy of the classification task. These parameters include number of clusters, *K*, number of bins for discretization purposes and the percentage of selected features. Based on the results obtained, the predictive accuracy can be improved by tweaking the number of bins and the percentage of features selected. The experiments conducted have also proven that the data transformation process for relational datasets, that transforms this data from a relational model representation into a vector space model representation, may produce unwanted features which may affect the accuracy of the predictive task. The optimum number of bins used for discretizing the numerical values in the dataset, cannot be justified as the highest predictive accuracies for all B1, B2 and B3 are obtained with different number of bins. This is due to the fact that, every feature that exists in the vector space model requires different number of bins because every feature has its own range of values due to the nature of the data transformation process performed by the DARA algorithm. This problem can be handled by applying the entropy based discretization method in discretizing each feature that exists in the record-pattern matrix. However, this method is very time consuming as the complexity of the entropy-based discretization depends also on the number of unique values.

References

1. Quinlan, J.R.: Learning Logical Definitions from Relations. Machine Learn. 5(3), 239–266 (1990)
2. Mitra, P., Murthy, C.A., Pal, S.K.: Unsupervised feature selection using feature similarity. Pattern Analysis and Machine Intelligence 24(3), 301–312 (2002)
3. Miller: Subset Selection in Regression, 2nd edn. Chapman & Hall (2002)
4. Quinlan, R.J.: C4.5: Programs for Machine Learning. Morgan Kaufmainn Series in Machine Learning (1993)

5. Srinivasan, A., Muggleton, S., Sternberg, M.J.E., King, R.D.: Theories for Mutagenicity: A Study in First-Order and Feature-Based Induction. Artificial Intelligence 85(1-2), 277–299 (1996)
6. Harigan, J.A.: Clustering Algorithms. John Wiley, New York (1775)
7. Witten, I.H., Frank, E.: Data Mining: Practical Machine Learning Tools and Techniques with Java Implementations. Morgan Kaufmann (1999)
8. Alfred, R.: The Study of Dynamic Aggregation of Relational Attributes on Relational Data Mining. In: Alhajj, R., Gao, H., Li, X., Li, J., Zaïane, O.R. (eds.) ADMA 2007. LNCS (LNAI), vol. 4632, pp. 214–226. Springer, Heidelberg (2007)
9. Alfred, R.: Optimizing feature construction process for dynamic aggregation of relational attributes. J. Comput. Sci. 5, 864–877 (2009), doi:10.3844/jcssp.2009.864.877
10. Alfred, R.: Summarizing relational data using semi-supervised genetic algorithm-based clustering techniques. Journal of Computer Science 6(7), 775–784 (2010)
11. Alfred, R.: Feature transformation: A genetic-based feature construction method for data summarization. Computational Intelligence 26(3), 337–357 (2010)
12. Salton, G., Michael, J.: McGill, Introduction to Modern Information Retrieval. McGraw-Hill Inc., New York (1986)
13. Karunaratne, T., Bostrom, H., Norinder, U.: Pre-Processing Structured Data for Standard Machine Learning Algorithms by Supervised Graph Propositionalization – a Case Study with Medicinal Chemistry Datasets. In: Ninth International Conference on Machine Learning and Applications, pp. 828–833 (2010)
14. Muggleton, SH.: Learning Stochastic Logic Programs, In Proceedings of the AAAI 2000 Workshop on Learning Statistical Models from Relational Data, Technical Report WS-00-06, pp.36-41 (2000)
15. Zhang, C., Wang, J.: Multi-relational Bayesian Classification Algorithm with Rough Set. In: 7th Intl. Conf. Of FSKD 2010, pp. 1565–1568 (2010)
16. Li, Y., Luan, L., Sheng, Y., Yuan, Y.: Multi-relational Classification Based on the Contribution of Tables. In: International Conference on Artificial Intelligence and Computational Intelligence, pp. 370–374 (2009)
17. Cao, P., Hong-yuan, W.: Multi-relational Classification on the Basic of the Attribute Reduction Twice. Communication and Computer 6(11), 49–52 (2009)
18. He, J., Liu, H., Hu, B., Du, X., Wang, P.: Selecting Effective Features and Relations For Efficient Multi-Relational Classification. Computational Intelligence 26(3), 1467–8640 (2010)
19. Guo, H., Herna, L.: Viktor: Multi-relational classification: a multiple view approach. Knowl. Inf. Systems 17, 287–312 (2008)
20. Wrobel, S.: Inductive Logic Programming for Knowledge Discovery in Databases: Relational Data Mining, pp. 74–101. Springer, Berlin (2001)
21. Emce, W., Wettschereck, D.: Relational instance-based learning. In: Proceedings of the Thirteenth International Conference on Machine Learning, pp. 122–130. Morgan Kaufmann, San Matco (1996)
22. Kirsten, M., Wrobel, S., Horvath, T.: Relational Distance Based Clustering. In: 8th International Conference on Inductive Logic Programming, pp.261–270 (1998, 2001)
23. Woznica, A., Kalousis, A., Hilorio, M.: Kernel-based distances for relational learning. In: Proceedings of the Workshop on Multi-Relational Data Mining at KDD (2004)
24. Getoor, L.: Multi-relational data mining using probalilistic relational models: research summary. In: Proceedings of the First Workshop in Multi-relational Data Mining (2001)

Generating Relevant and Diverse Query Suggestions Using Sparse Manifold Ranking with Sink Regions

Van Thanh Nguyen and Kim Anh Nguyen

School of Information and Communication Technology
Hanoi University of Science and Technology
thanhng.cs@gmail.com, anhnk@soict.hut.edu.vn

Abstract. In order to improve the usability of a search engines, Query Suggestion, a technique for generating alternative queries to Web users, has become an indispensable feature for such systems. By measuring the similarity between queries in the Euclidean space, however, most existing works mainly focus on suggesting relevant queries to the original query while ignoring diversity in the suggestions, which will potentially dissatisfy Web users' information needs. In fact, it is more natural and reasonable to assume that the query space is a sparse manifold. In this paper, we present a novel query suggestion method based on sparse query manifold learning and sparse manifold ranking with sink regions. By turning selected queries and their sparse neighbors into sink regions on the sparse query manifold, our approach can extract query suggestions by simultaneously considering both diversity and relevance in a unified way. Empirical experimental results on a large scale query log show that our approach is able to effectively generate highly diverse as well as semantically related suggestions.

1 Introduction

With the advent of search engines, it is increasingly easier for Web users to seek desired information. In order to enhance the effectiveness, Query Suggestion has long been proved indispensable to help users explore and express their search intents. Most previous methods mainly focus on providing users with the most relevant results based on some well-defined criteria. However, for query suggestion, only providing the "good" suggestions is separate from satisfying users' information needs. Actually, users tend to submit short queries containing one or two ambiguous terms which may make these approaches confused with producing suggestions [1]. Furthermore, in most cases, they cannot clearly express their purposes in several query words due to the lack of knowledge. In order to find satisfactory answers, the users have to rephrase their queries constantly [2]. Therefore, the suggestions should be semantically different from each other in order to cover many possible aspects of users' intent. In addition, previous query suggestion approaches mostly relied on measuring the similarity between queries in the Euclidean space, either based on query terms or click-through data. Nevertheless, there is no convincing evidence that the query space is Euclidean. It is more natural to assume that the query space is a manifold, either linear or nonlinear. The local geometric structure is essential to reveal the relationship between

A. Selamat et al. (Eds.): ACIIDS 2013, Part I, LNAI 7802, pp. 176–185, 2013.

queries. Moreover, ranking queries in terms of the intrinsic global manifold structure is superior to the pair-wise distance in the Euclidean space [3].

To model the query manifold structure, a proper choice of the neighborhood size used to build the neighborhood graph is critical. Moreover, in some cases one cannot find a neighborhood that contains only points from the same manifold, especially when manifolds are close to each other. Relating to this problem, E. Elhamifar et al. [4] propose an optimization program based on sparse representation to select a few neighbors of each data point that span a low-dimensional affine subspace passing near that point. Inspired by the research work on sparse manifold embedding [4], we find that it is more reasonable to consider that the query space is a sparse manifold. As a result, a few nonzero elements of a sparse solution indicate sparse neighbors of query that are on the same manifold. This leads to successful results even in challenging situations where the nearest neighbors of a query come from other manifolds.

In this paper, we propose a novel method to recommend relevant and diverse queries, named sparse manifold ranking with sink regions. It leverages a ranking process over sparse query manifold which can naturally make full use of the relationships among queries to find relevant and salient queries. By turning selected queries and their sparse neighbors into sink regions on the sparse query manifold, our approach can generate suggestions by simultaneously considering both diversity and relevance. We conducted extensive experiments to evaluate the proposed method based on a large collection of query logs from a commercial search engine. Empirical experimental results demonstrate the effectiveness of our query suggestion method.

The main contributions of our work can be summarized as follows: (1) we first learn the sparse query manifold structure to discover the local geometry structure of queries; (2) we propose a novel ranking approach, i.e., sparse manifold ranking with sink regions for query suggestions which addresses relevance and diversity simultaneously in a unified way; (3) we show that our method outperforms four other baselines in generating relevant and diverse query suggestions. The rest of the paper is organized as follows. Section 2 introduces related work. Section 3 describes our proposed approach in detail. Experimental results are discussed in Section 4 and conclusion is made in Section 5.

2 Related Work

Query Suggestion. Improving the performance and quality of query suggestion techniques has been extensively studied in the past decades to enhance the entire user search experience within the same search intent. Query suggestion approaches [1, 2, 5, 6] mostly relied on measuring the similarity between queries in the Euclidean space, either based on query terms or click-through data. However, most previous work only took relevance of suggestions into account, while not explicitly addressed the problem of diversity. Mei et al. [2] tackled this problem using hitting time on the query-URL bipartite graph. Although their approach can suggest more diverse queries by boosting long tail queries, the main weakness of their approach is that it would sacrifice the relevance considerably when improving diversity, and many long tail queries recommended to users may not be familiar to them.

Sparse Manifold Learning. In many real-world problems, the data lie in multiple manifolds of possibly different dimensions. The construction of the neighborhood is a critical problem of manifold learning. The choice of a neighborhood size is a difficult task and only weak heuristics can be stated. Since curvature and density may vary over the manifold, one global setting may not work well for the entire manifold. Based on the geometrically motivated assumption that for each data point there exists a small neighborhood in which only the points that come from the same manifold lie approximately in a low-dimensional affine subspace, authors in [4] proposed an optimization program to select a few neighbors of each data point that are on the same manifold. In addition, the weights associated to the chosen neighbors indicate their distances to the given data point. Thus, their method simultaneously builds the neighborhood graph and obtains its weights.

The closest work to ours is a unified framework for recommending diverse and relevant proposed by Zhu et al. [7]. The authors built a *k-nearest-neighbor query graph* using the click-through information in query logs to model the local query manifold structure and apply a manifold ranking process over query manifold to find relevant and salient queries. In order to achieve diversity, it turns the ranked queries into stop points on the query manifold that stop spreading their ranking scores to their nearby neighbors during the manifold ranking process. Although the *k-NN query graph* can guarantee the sparseness of the query structures in query logs, there are a number of problems associated with the choice of the stable parameter *k*. A large number of nearest neighbors easily makes them fail to nicely deal with the highly twisted and folded manifold and so, some neighbors of one query can come from other different folds [8]. In contrast, too small a neighborhood can falsely estimate the relationships between the queries and their neighbors, even divide the continuous manifold into disjoint sub-manifolds. Moreover, by turning only selected queries into stop points, some nearest neighbors of selected queries still can receive a high rank during the manifold ranking process that can decrease diversity of suggestions.

Different from this approach, we first learn the sparse query manifold structure to measure the similarity between queries using an optimization program based on sparse representation. Then, we propose a novel sparse manifold ranking with sink regions for query suggestion. By turning selected queries and their sparse neighbors into sink regions, it can effectively prevent semantically redundant queries from receiving a high rank, hence encouraging diversity in the results. Therefore, our approach can generate query suggestions by simultaneously considering both diversity and relevance in a unified way. Like traditional manifold ranking algorithm, the new proposed ranking approach also shows a nice convergence property.

3 Proposed Approach

3.1 Sparse Query Manifold Learning

Assume we are given a collection of N data points (i.e. queries) $\{x_i \in R^D\}^N_{i=1}$ lying in n different manifolds $\{M_l\}^n_{l=1}$ of intrinsic dimensions $\{d_l\}^n_{l=1}$. We need to build a similarity graph whose nodes represent the data points and whose weighted edges represent the similarity between data points. To do sparse query manifold learning, we wish to connect each point to other points from the same manifold with appropriate weights that reflect the neighborhood information. This problem is formulated as an optimization algorithm based on sparse representation [4]. The assumption behind the proposed method is that each data point has a small neighborhood in which the minimum number of points that span a low-dimensional affine subspace passing near that point is given by the points from the same manifold.

Formally, we consider a point x_i in the d_l-dimensional manifold M_l and the set of points $\{x_j\}_{j \neq i}$. It follows from above assumption that, among these points, the ones that are neighbors of x_i in M_l span a d_l-dimensional affine subspace of R^D that passes near x_i. In other words,

$$\|[x_1 - x_i \ldots x_N - x_i]\, c_i\|_2 \leq \varepsilon \text{ and } 1^{\mathrm{T}} c_i = 1 \tag{1}$$

has a solution c_i whose $d_l + 1$ nonzero entries correspond to $d_l + 1$neighbors of x_i in M_l. Because the solution c_i with the smallest number of nonzero entries may not be unique, we use an optimization program whose objective function favors selecting a few neighbors of x_i subject to the constraint in (1), which enforces selecting points that approximately lie in an affine subspace at x_i. To do so, we normalize the vectors $\{x_j - x_i\}_{j \neq i}$ and let $X_i = [X_{i1}, X_{i2}, \ldots, X_{iN}]$ where $X_{ij} = \frac{x_j - x_i}{\|x_j - x_i\|_2}, j \neq i.$

Now, among all the solutions of $\|X_i c_i\| \leq \varepsilon$ that satisfy $1^{\mathrm{T}} c_i = 1$, we look for the one that uses a few closest neighbors of x_i. We thus consider the following weighted P_1-optimization program where the P_1-norm promotes sparsity of the solution [4, 9] and the proximity inducing matrix Q_i, which is a positive-definite diagonal matrix, favors selecting points that are close to x_i

$$\min \|Q_i c_i\|_1 \text{ subject to } \|X_i c_i\| \leq \varepsilon,\ 1^{\mathrm{T}} c_i = 1 \tag{2}$$

A simple choice of the proximity inducing matrix is to select the diagonal elements of Q_i to be $\frac{\| x_j - x_i \|_2}{\sum_{t \neq i} \| x_t - x_i \|_2} \in (0, 1]$.

By the method of Lagrange multipliers, we have to find the sparse solution of the following convex optimization problem

$$\operatorname{argmin} \tfrac{1}{2} \| X_i c_i \|^2_2 + \lambda \|Q_i c_i\|_1 + \beta 1^T c_i \tag{3}$$

This problem is similar to the Lasso optimization ones. Therefore, from the Krush-Kuhn-Tucker optimality conditions, a necessary and sufficient condition for (3) is:

$$\begin{gathered} X_{ij}^T (X_i \hat{c}_i) + \frac{\lambda Q_{ij} \hat{c}_{ij}}{\| \hat{c}_{ij} \|} + \beta = 0, \forall c_{ij} \neq 0 \\ \| X_{ij}^T (X_i \hat{c}_i) + \beta \| \leq \lambda Q_{ij}, \forall c_{ij} = 0 \end{gathered} \tag{4}$$

Based on this stationary condition, an iterative blockwise coordinate descent algorithm [9] can be derived, a solution to (4) satisfies

$$\hat{c}_{ij} = \left[1 - \frac{\lambda Q_{ij}}{\| S_{ij} \|} \right]_+ S_{ij} \tag{5}$$

Where $S_{ij} = X_{ij}^T (X_i c_{\backslash ij}) + \beta$, with $c_{\backslash ij} = (c_{i1}, \ldots, c_{ij-1}, 0, c_{ij+1}, \ldots, c_{iN})^T$. By iteratively applying (5), the sparse solution can be obtained.

The solution c_i of the proposed optimization programs satisfies $\sum c_{ij}X_{ij} \approx 0$. Hence, we can rewrite $x_i \approx [x_1x_2...x_N]\omega_i$ where the weight vector $\omega_i^T = [\omega_{i1}...\omega_{iN}] \in R^N$ associated to the i-th data point is defined as

$$\omega_{ii} \overset{def}{=} 0, \omega_{ij} \overset{def}{=} \frac{c_{ij} / \|x_j - x_i\|_2}{\sum_{t \neq i} c_{it} / \|x_t - x_i\|_2}, \forall j \neq i \tag{6}$$

The indices of the few nonzero elements of ω_j, ideally, correspond to neighbors of x_i in the same manifold and their values indicate their inverse distances to x_i. Because of the sparsity of solution c_i, each point x_i has only a few closest neighbors in the same manifold. We call such neighbors as sparse neighbors of x_i.

We define an affinity matrix W for the sparse query manifold where $W_{ij} = (\omega_{ij}+\omega_{ji})/2$. The computational complexity of the iterative algorithm for solving the optimization program (3) is $O(kDN^2)$ where k is the number of iterations. Hence, its computational time increases considerably as N increases. However, the sparse query manifold learning process is conducted offline and periodically. In addition, in order to find a solution for each query point, we consider only queries that have at least one common clicked URL with that point. Therefore, the convex optimization program reaches a reasonable convergence within a few iterations.

3.2 Sparse Manifold Ranking with Sink Regions

Given a set of data points (i.e. queries) $X = \{x_0, x_1 \ldots x_N\} \subset \mathrm{R}^D$, the first point x_0 is the input query and the rest of the points x_i $(1 \leq i \leq N)$ are the candidate queries. Let f: $X \rightarrow R$ denotes a ranking function which assigns to each point x_i $(0 \leq i \leq N)$ a ranking value f_i. We can consider f as a vector $f = [f_0, \ldots , f_N]^T$. We also define a vector $y = [y_0,..., y_N]^T$, in which $y_0 = 1$ for the input query x_0 and $y_i = 0$ for the others.

A traditional ranking process [3] over the query manifold is described as follows:

1. Symmetrically normalize W by $S = D^{-1/2}WD^{-1/2}$ in which D is the diagonal matrix with (i, i)-element equal to the sum of the i-th row of W.
2. Iterate $f(t+1) = \alpha Sf(t) + (1 - \alpha)y$ until convergence, where $\alpha \in [0, 1)$.
3. Let f^*_i denote the limit of the sequence of $\{f_i^{(t}\}$. Rank each point q_i according its ranking scores f^*_i (largest ranked first).

In the above ranking process, all the points spread their ranking scores to their neighbors via the weighted edges. The spread process is repeated until a global stable state is achieved and all the points are ranked according to their final scores. With the traditional manifold ranking process, we can obtain relevant and salient queries for suggestion given an input query.

Moreover, to capture the diversity during the ranking process, because the sparse neighbors of a query are on the same manifold, we turn selected queries and their sparse neighbors into stop points on the sparse query manifold that stop spreading their ranking scores to their nearby neighbors during the manifold ranking process. The ranking scores of other queries close to these stop points (i.e., queries which share similar search intent with the selected queries) will be naturally penalized during the ranking process over the sparse manifold. A selected query and its sparse neighbors are called a sink region. Therefore, our approach can generate query suggestions by simultaneously considering both diversity and relevance in a unified way. Here we derive the new iteration algorithm for sparse manifold ranking with sink regions. Let *T* denote the set of points in sink regions, and *R* denote the set of free points (all the data points excluding the points in sink regions). The normalized matrix *S* in traditional manifold ranking can then be reorganized as a block matrix $\begin{pmatrix} S_{RR} & S_{RT} \\ S_{TR} & S_{TT} \end{pmatrix}$. Since points in sink regions never spread their scores to points to which they are connected, we set $S_{RT} = S_{TT} = 0$, then we get the new iteration equation for sparse manifold ranking with sink regions:

$$f_R^{(t+1)} = \alpha S_{RR} f_R^{(t)} + (1 - \alpha) y_R \tag{7}$$

where f_R denote the ranking scores of points in set *R* and y_R denote the prior on the points in set *R* and the parameter *α* specifies the relative contributions to the ranking scores from neighbors and the initial ranking scores.

3.3 Generating Relevant and Diverse Query Suggestions

Our query suggestion method is finally obtained as follows. We first do sparse manifold learning to find a few neighbors for each query point and then model the query manifold structure and set all the query points as free points. Giving an input query, we apply the proposed ranking algorithm, i.e., sparse manifold ranking with sink regions, over the query manifold until a global stable state is achieved, and rank the queries according to their ranking scores. The free point (exclusive of the input query) with the largest ranking score will be selected as a suggestion. That query and its sparse neighbors (except the input) are then turned into a sink region in the posterior iteration. In the online part, the suggestion algorithm using sparse manifold ranking with sink regions is shown in Algorithm 1.

Algorithm 1. Query Suggestion using Sparse Manifold Ranking with Sink Regions

```
Input: $q$ - input query                 $\chi$ - all the other queries
K - suggestion size                      $S$ - normalized matrix of the sparse manifold
$T$ - set of query points in the sink regions      $R$ - free point set
Output: Top K suggestion query in the set $U$
Initialization: $U = \phi$, $T = \phi$, $R = \chi$
for k = 1..K do
    Obtain $S_{RR}$ based on $S$, $T$ and $R$.
    Iterate $f_R^{(t+1)} = \alpha S_{RR} f_R^{(t)} + (1-\alpha) y_R$ until convergence obtained with $f_R^{(0)} = 0$ where $\alpha \in [0,1)$.
    Select the query $q_k$ with the largest ranking score (except the input query) as a suggestion $U = U \cup \{q_k\}$.
    Turn the selected queries and their sparse neighbors into sink regions, $T = T \cup \{q_k \cup N(q_k)\}$ and $R = R \setminus \{q_k \cup N(q_k)\}$, where $N(q_k)$ – sparse neighbors of $q_k$ except $q$
End for
```

Note that the above ranking algorithm still has a nice convergence property with the introduction of sink regions. Besides, in the presence of these sink regions, the complexity of the ranking algorithms is significantly reduced. Moreover, most queries are irrelevant to the input ones, we can use a width first search strategy to construct a sub-manifold to save the computational cost.

4 Experiments

4.1 Data Set

Our experiments are based on the click-through data of American Online query log published in 2006 which contains about 20M queries. We utilize a sample of 2.2M records and extract queries along with clicked URLs. To reduce the noise in the data set, we cleaned the raw data and discarded queries and URLs with a frequency less than 4. After cleaning, we obtained the click-through data with totally 114,481 unique queries and 210,053 unique URLs. On average, each query clicks 1.84 URLs and each URL is clicked by 2.76 distinct queries. We randomly sampled 100 queries with frequencies between 100 and 2,000 for comparison.

4.2 Evaluation Criteria and Baselines

Criteria. Evaluating the quality of query recommendation is difficult, since there is usually no ground truth of recommendations. Besides, there is no evaluation metric that seems to be universally accepted as the best for measuring the performance of recommendation algorithms. Therefore, we adopt the following three metrics (*Relevance*, *Diversity* and *Q-measure*) used in [7] to help evaluate the relevance and diversity in suggestions.

Baselines. We demonstrate the performance of our approach, called Sparse_Mani for short, with four following baselines:

- *Naïve*: Each query is represented as a URL vector and use Euclidean distance between queries as a measurement. For a given query q, queries with smallest distance scores are selected for suggestions.
- *MMR* (Maximal Marginal Relevance) [10]: MMR measures the relevance and diversity independently and provides a linear combination, called "marginal relevance", as the metric. Formally,

$$MMR \triangleq Arg \min_{q_i \in R \setminus S} [\lambda dist(q_i, q) - (1-\lambda) \min_{q_j \in S} dist(q_i, q_j)] \tag{8}$$

  where R is a set of candidate queries, S is the subset of queries in R which is already selected, *dist* is the distance metric between queries and λ is a parameter for linear combination. For a given query q, MMR will iteratively recommends queries with the largest "marginal relevance".
- *Manifold ranking with stop point*, Mani_stop for short, leverages ranking process over query manifold to find relevant and salient suggestions. Meanwhile, the concept "stop point" was introduced on the manifold to address the diversity.

Stop points naturally penalize ranking scores of queries close to them. To model the query manifold structure, in our experiments, we construct *k-nearest-neighbor graph* and *ε-graph*.

4.3 Parameter Settings

By testing over a variety of parameter, we apply the following parameter in our experiments. For MMR, we selected the best λ = 0.6. We also tested and chose k = 50 and *ε = 1.28* (distance between queries) to model query manifold structure for Mani_stop. In our approach, we typically set the sparsity trade-off λ = 100. Finally, we fixed the parameter α = 0.99 in our method (Sparse_Mani) as well as in Mani_stop.

4.4 Results and Discussion

Our experiments were carried over the numbers of suggestions associated with each query and the results are presented in Figures 1-3. We choose only 6 suggestions for each query because of the sparseness. In addition, a large suggestion size may be unnecessary and even decrease the suggestion quality.

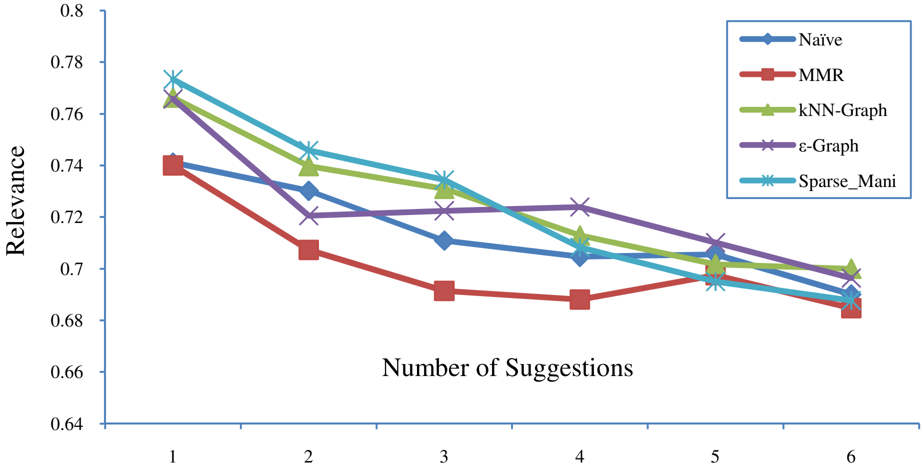


Fig. 1. Average Relevance over the Suggestion Size under five Approaches

Figure 1 shows the average relevance values of query suggestions under five different methods. By nature, the average relevance value gradually decreases when the suggestion size increases. We notice that Sparse_Mani and Mani_stop including *kNN-graph* and *ε-graph* achieve higher performance in relevance as compared with MMR and Naive methods which directly measure the relevance relied on Euclidean distance. Moreover, the relevance of these two approaches lessens more quickly because of taking diversity into account. It demonstrates the effectiveness of the intrinsic query manifold in capturing the relevance between queries. Besides, we can see that the average relevance value of our method is better than of the others of Mani_stop. This is because of asymmetry of solving optimization problem for each query and we always obtain higher weights for closer queries.

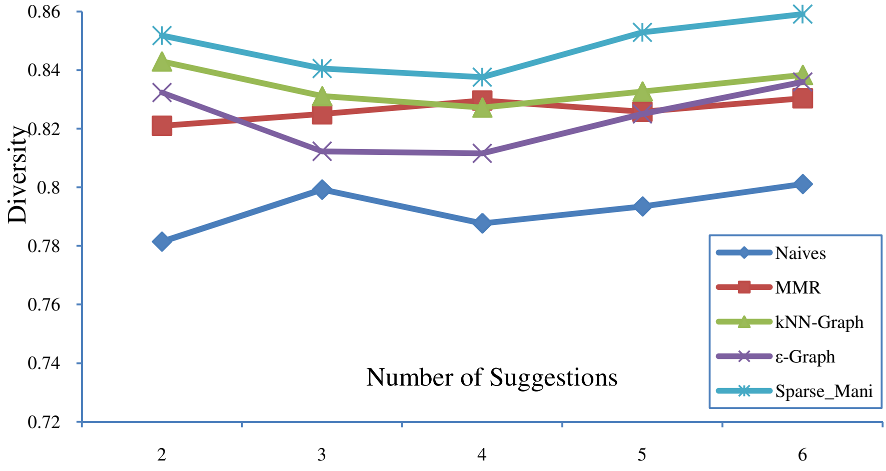


Fig. 2. Average Diversity over the Suggestion Size under Five Approaches

The average diversity values of query suggestions under the five different approaches are shown in Figure 2. Not surprisingly, the diversity of Naive is the lowest one in the five approaches, since Naive only focuses on suggesting queries according to their relevance with the input query. Both MMR, Sparse_Mani and Mani stop can receive higher diversity value by explicitly address the diversity in ranking. Among these three approaches, Sparse_Mani obtains the highest diversity on average by introducing the sink regions into query manifold. It indicated the coherency and the efficiency of the "sink regions" over the sparse query manifold.

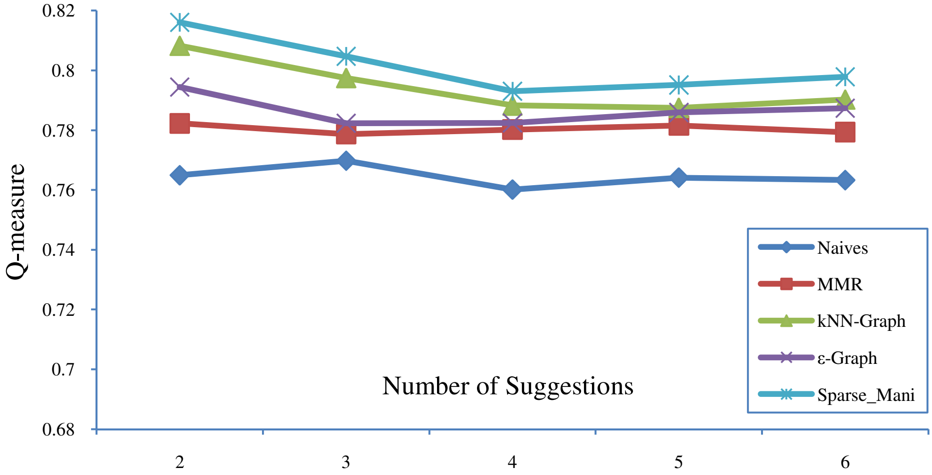


Fig. 3. Q-measure over the Suggestion Size under Five Approaches

To better evaluate the overall effectiveness of different approaches, we show the Q-measure in Figure 3. Among the five approaches, the proposed Sparse_Mani approach consistently outperforms the other four other baselines in terms of Q-measure. These above results clearly aver that our approach can present with significant improvement in generating diverse and relevant suggestions.

5 Conclusions

This paper presents a novel unified query suggestion method based on sparse query manifold learning and sparse manifold ranking with sink regions. By applying the sparse representation program, we learn the local structure for each query and automatically obtain similarity between queries. Besides, by turning selected queries and their sparse neighbors into sink regions on the sparse query manifold, our approach can extract query suggestions by simultaneously considering both diversity and relevance in a unified way. Empirical experimental results over three standard evaluation criteria demonstrate that our approach outperforms the four baselines.

For the future work, we would like to make full use of the dimensional information of the manifolds and apply the proposed approach to a variety of problems such as text summarization, text document clustering and classification.

References

1. Cao, H., Jiang, D., Pei, J., He, Q., Liao, Z., Chen, E., Li, H.: Context-aware query suggestion by mining click-through and session data. In: Proceeding of th 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 875–883 (2008)
2. Mei, Q., Zhou, D., Church, K.: Query suggestion using hitting time. In: Proceeding of the 17th ACM Conference on Information and Knowledge Management, pp. 469–477 (2008)
3. Zhou, D., Weston, J., Gretton, A., Bousquet, O., Scholkopf, B.: Ranking on data manifolds. In: Proceedings of the 17th Annual Conference on Neural Information Processing Systems (2003)
4. Elhamifar, E., Vidal, R.: Sparse manifold clustering and embedding. In: NIPS (2011D)
5. Beeferman, D., Berger, A.: Agglomerative clustering of a search engine query log. In: Proceedings of the Sixth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 407–416 (2000)
6. Ma, H., Yang, H., King, I., Lyu, M.R.: Learning latent semantic relations from click-through data for query suggestion. In: Proceeding of the 17th ACM Conference on Information and Knowledge Management, pp. 709–718 (2008)
7. Zhu, X., Guo, J., Cheng, X., et al.: A unified framework for recommending diverse and relevant queries. In: WWW 2011, India (2011)
8. Wen, G.: Relative transformation-based neighborhood optimization for isometric embedding. Neurocomputing 72, 1205–1213 (2009)
9. Yuan, M., Lin, Y.: Model selection and estimation in regression with grouped variables. Journal of the Royal Statistical Society, Series B 68, 49–67 (2006)
10. Carbonell, J., Goldstein, J.: The use of mmr, diversity-based reranking for reordering documents and producing summaries. In: Proceedings of SIGIR 1998, New York, USA, pp. 335–336 (1998)

Measuring Data Completeness for Microbial Genomics Database

Nurul A. Emran[1], Suzanne Embury[2], Paolo Missier[3],
Mohd Noor Mat Isa[4], and Azah Kamilah Muda[1]

[1] Centre of Advanced Computing Technology (C-ACT), University Teknikal Malaysia Melaka
[2] The University of Manchester
[3] The University of Newcastle
[4] Genome Malaysia Institute
{nurulakmar,azah}@utem.edu.my, embury@cs.man.ac.uk,
paolo.missier@ncl.ac.uk, emno72@gmail.com

Abstract. Poor quality data such as data with missing values (or records) cause negative consequences in many application domains. An important aspect of data quality is completeness. One problem in data completeness is the problem of missing individuals in data sets. Within a data set, the individuals refer to the real world entities whose information is recorded. So far, in completeness studies however, there has been little discussion about how missing individuals are assessed. In this paper, we propose the notion of population-based completeness (PBC) that deals with the missing individuals problem, with the aim of investigating what is required to measure PBC and to identify what is needed to support PBC measurements in practice. This paper explores the need of PBC in the microbial genomics where real sample data sets retrieved from a microbial database called Comprehensive Microbial Resources are used (CMR)[1].

Keywords: data completeness, population-based completeness (PBC), completeness measurement.

1 Introduction

One type of completeness that has been mentioned in the literature is population-based completeness (PBC) [1]. A population consists of a set of individuals that represent real world entities, whose information are recorded in a data set. For PBC, the concern is to determine whether the data set consists of a complete set of individuals or not, which requires measuring the individuals that are missing relative to a population. The importance of PBC can be seen in the descriptions of many problems in the literature. For example, in bioinformatics, to study the genes that are responsible for certain diseases, a candidate gene set is prepared and validated before more detailed tests are performed to find the disease-causing

[1] CMR-http://www.tigr.org/CMR

A. Selamat et al. (Eds.): ACIIDS 2013, Part I, LNAI 7802, pp. 186–195, 2013.

genes [2]. According to Tiffin *et al.*, because many complex diseases could be linked to multiple gene combinations, determining whether the data set of gene candidates is complete or not is becoming more important in the analysis for the bioinformaticians in order to produce a more reliable set of (potential) disease-causing genes [3]. Consequently, if some genes are missing from the gene data set used in the analysis, links between those genes and the disease cannot be established.

In this example, completeness of the candidate gene data set used in the analysis is determined by consulting various gene data sources like public genome databases, gene expression databases, data on gene regulatory networks and pathways, as well as biomedical literature to check whether any gene has been missed from the data set [3]; the genes reference population is gathered from multiple sources. The need for PBC is not limited to the example just mentioned where the usage of reference populations is crucial to determine completeness of data sets under measure. However, little is known about how the reference populations are defined. In addition, how PBC is measured is unclear as the measurement method(s) used has not been described formally. Ideally, the reference populations used in PBC measurements are the representation of the real world which we would call the *true populations*. However, using true populations for PBC measurements can be hindered by the lack of knowledge of the individuals of the true populations or by the inaccessibility of the source(s) that provides information about the individuals. The alternatives to true populations are 'approximate' populations, the populations that could be accepted by the application domain's community as 'complete' reference populations in PBC measurements. However, even when approximate populations are adopted, we cannot avoid answering fundamental questions of PBC that unfortunately have not been addressed by any studies in completeness.

The rest of this paper is organised as follows. Section 2 covers the various types of data completeness proposed to date; Section 3 consists of the elements essential for PBC, Section 4 presents the example of PBC. Finally Section 5 concludes the paper.

2 Related Work

Studies in data completeness are not new; they have been conducted since at least the 1970s. During this period, the data completeness problem was well known as the problem of missing information among scholars in the database community [4] as well as among statisticians [5]. For the database community, the early work on completeness largely dealt with the problem of representing missing values (as opposed to 'empty' or undefined values) within the relational tables, where *nulls* were usually assigned for the missing values in the tables [6]. Various representations of null have been used, for example, the @ symbol [7], ω [4] and the use of variables such as x, y and z [8]. The first proposal for a measure of null-based completeness (NBC) was made by Fox, Levitin and Redman [9]; they described a datum as a triple $< e, a, v >$, where v is the value of the attribute a that belongs to an entity e [9]. Nulls were viewed from two levels of granularity: single datum level and at data collection level. At the single datum level, a *binary measure* was proposed which checks whether a datum has a value or not; at the collection level, the study described the completeness measure as an *'aggregate' measure* that computes the fraction of the data that are null.

The tuple-based completeness (TBC) measure proposed by Motro and Rakov [10] is not only useful for detecting missing tuples, but it also helps to determine whether the tuples are accurate. TBC, in their proposal was viewed from a database level and is defined as an 'aggregate' measure as follows:

$$Completeness(of\, the\, database\, relative\, to\, the\, real\, world) = \frac{|D \cap W|}{|W|},$$

where D is the actual stored database instance while W is the ideal, real world database instance. From this definition, we gain an important insight into completeness which is completeness can be affected by the presence of errors in the data set. W in the definition represents not only a reference data set that is complete, but also a reference data set that is accurate. Nevertheless, because W is very unlikely to be acquired, the measure used the sample of W which came from alternative databases or judicious sampling (where the verification of the samples is made by humans) [11].

Schema-based completeness (SBC) however focuses on "model completeness" where Sampaio and Sampaio defined it as "the measure of how appropriate the schema of the database is for a particular application". From an XML point of view, Sampaio and Sampaio defined SBC as the number of missing attributes relative to the total number of attributes [12].

To the best of our knowledge, the first recorded use of the term 'population' in connection with completeness is in a proposal by Pipino, Lee and Wang [1]. The authors did not provide a formal definition of the PBC measure, but hinted at the presence of this useful concept through an example. In the example, the authors stated that, "If a column should contain at least one occurrence of all 50 states, but only contains 43 states, then we have population incompleteness" [1]. From the example, we observe that there is a data set under measure (from state column) in which its completeness is determined by the number of missing 'individuals' (the states) from a 'reference population' (a set of 50 states). There is a notion of reference population that is used to represent a population that consists of complete individuals. However, details of how PBC measurement is made in practice are missing from both proposals, especially in terms of how the reference populations are acquired and used. The elaboration of the concept of PBC therefore remains an open question for research in terms of the current literature. To continue exploring the notion of PBC, we present the elements essential to measure PBC in the next section.

3 The Elements of PBC

The examples that hinted at PBC given by Pipino, Lee and Wang [1] and by Scannapieco and Batini [13] help us to understand that the authors have a similar concern to each other, which is on the 'individuals' that are missing from a population. They help us to see that completeness is not only about counting nulls or missing tuples in data sets which receives the most literature coverage. We observe from the examples that, to measure PBC, we need data sets to be measured and 'reference' populations. An explanation of how data sets under measure and their reference populations are used in terms of a formal measurement definition is, however, missing from the literature.

As presented Section 2, Motro and Rakov proposed a TBC measure [10], where the formal definition of the measure is as a *simple ratio method.* We apply a similar form of simple ratio method in our PBC measure and define a basic PBC measurement as:

$$Completeness(D, RP) = \frac{|(D \cap RP)|}{|RP|} \in [0, 1], \tag{1}$$

where D is the data set under measure, and RP is the reference population.

We can see from Equation (1) that measuring PBC is conceptually simple as we only need a data set to measure and a reference population. Nevertheless, to make the measurement workable in practice, we need to know more about the populations. The question of how the reference populations can be acquired is also essential, especially in the context of PBC measurement providers.

3.1 Populations

The term *population* is used widely, especially in statistical studies. Statisticians define a population as the entire collection of items that form the subject of a study and that share common features [14]. Within the statistical studies themselves, the definition of a population however is often specific to the application domain. For example, statistical studies in the biological and ecological domains define a population as a group of organisms of the same species in a given area ([15]). In census studies, a population is defined as the people who inhabit a territory or a state [16]. These items, species or people are the 'individuals' that belong to their defined population. In philosophy, the term *natural kind* is used for "grouping or ordering that does not depend on humans", which is the opposite for the term *artificial kind* used for grouping of arbitrary things made by human [17]. Inspired from the observation of how populations are defined in the literature and from the philosophical domain, we define population as *a set of individuals of a natural kind* and these individuals are the real world individuals (not the artificial individuals created by humans). A question that arises is: what characterises the individuals that are suitable to act as the members of populations for PBC?

Pipino, Lee and Wang pointed out in their example that, the data set that they examined are retrieved from a specific column (states) [1]. This provides us with a hint that only certain attribute of a data set might be of interest and will 'make sense' as the basis of a completeness measure. The instances in state column are therefore the data set under measure that is of interest in terms of its completeness. Thus in the example, the individuals that are suitable to act as the members of a population are a set of states.

3.2 Reference Populations

The notion of reference populations is proposed as an essential element of PBC to represent populations that are 'complete', i.e., that have no missing individuals. The question is, how can we obtain the reference population? In bioinformatics completeness of the human gene population that is used for an analysis for genes that cause diseases is of concerned [3]. Several gene sources such as public genome databases, gene expression databases, data on gene regulatory networks and pathways, as well as biomedical literature were used to retrieve the list of genes that became the reference population [3]. The reference population used in this study is the integration of several sets of genes from a variety of gene sources (identified by the bioinformaticians in the domain).

However, the reference population used in the analysis is not the *true* human gene population unless it consists of all real world human genes that exist. Because there is still a debate on the actual number of human genes among the scientists, and more time is required to discover true human genes, due to the complexity of the gene discovery process [18], the usage of the true gene population is not possible in this example. Therefore, as the alternative, we must find an approximate population to represent the true gene population.

Two forms of reference populations are possible: 1) the *true* populations that consist of all real world individuals that exist, and, 2) the *approximate* populations that are used to represent the true populations and which are more easily available. In addition, we also observe that, the individuals of a reference population may come from multiple sources. Within an application domain, we say that the decision regarding which form of reference population is to be used must be made by the domain experts (e.g., by bioinformaticians) due to their knowledge of the *sources* of the populations. The decision to use approximate populations in the examples above is driven by the costs/difficulties of acquiring the true populations, and the questionable benefits of the small differences in measurements that would result. If this is the case, the approximate populations used must be adequate for determining completeness of data sets within the domain. However, we suspect that the main reason approximate populations are used is that true populations are not feasible to obtain, even though there may be a need to use them.

The situation where there exists a single source that contains a good approximation of the true population is limited however (an exception being the Genbank database[2] that contains the genes with good evidence of their existence). We propose that good approximate populations should be established by integrating individuals from a range of sources. To describe approximate populations established from integrated sources, we adopt the term *universe of discourse (UoD)* or in short *universe*[3]. Conceptually, a universe consists of a

[2] http://www.ncbi.nlm.nih.gov/genbank/

[3] The term *universe* was introduced by Augustus De Morgan in 1846 in formal logics to represent the collection of objects under discussion of a specific discourse [19].

collection of approximate populations within an application domain used for PBC measurements, that is built by integrating individuals from several incomplete sources for the populations.

We use the term *contributing sources* (CSs) for sources that contribute to the reference populations in the universe. The CSs could be in multiple forms, such as databases (private and public) e.g., observation databases from gene regulatory networks and pathways [3] or published literature. As it is crucial to understand (and to manage) the relationship between the CSs and the reference populations for successful integration, we propose a structure called a *population map*. Conceptually, a population map consists of a mapping between a reference population and its CSs. If a reference population is stored as a table, and the CSs are databases, we say that a population map is a mapping between a reference population table and queries over tables on CSs. Note that, as the schema of the universe may not be the same as the schema of the CSs, the designers of the population maps must consider the differences.

4 Example of PBC Measurements in Microbial Genomics

The study of the genomes of microbes, called microbial genomics, helps pharmaceutical researchers gain a better understanding of how pathogens cause disease [20]. By understanding the association between pathogens and diseases, further analyses, such as regarding pathogens' resistance to drugs or antibiotics, can be conducted in search of a cure for specific diseases.

To explain the PBC problems in the microbial domain, we observed the relationships among the subjects of microbial studies that have been documented in the literature and we produced Fig. 1 as the result of these observations. The left side of the figure depicts an ER diagram with five subjects of microbial studies, namely *Microbe, Genome sequence, Gene, Infectious disease* and *Antimicrobial agent/vaccine*, and their relationships. Each relationship between the subjects is related to an analysis within the pathways of microbial genomics (the right side of the figure), shown by a dotted line. The analyses are conducted by the scientists in the wet lab through experiments, or by the bioinformaticians in the dry lab with the support of computational tools [2,21].

We observe that for every analysis in the microbial study pathway shown in Fig. 1, the scientists/bioinformaticians need to prepare an input data set describing the subjects of interest (e.g., *microbe and gene*) for the analysis. In general, in these analyses, the completeness of the input data set determines the completeness of the analysis result. Therefore, an important question that arises in this domain is regarding the completeness of these input data sets. However, not all information in these input data sets are of interest (in terms of completeness) as scientists often look at specific information that is important to them (i.e., completeness in regards to certain genes or species) - a scenario that hinted PBC problems that are inherent in the multiple stages of analysis in microbial domain as described. The key lesson that we learnt based on the observation in microbial domain is on the applicability of the PBC concept to support answering PBC measurement requests from this domain.

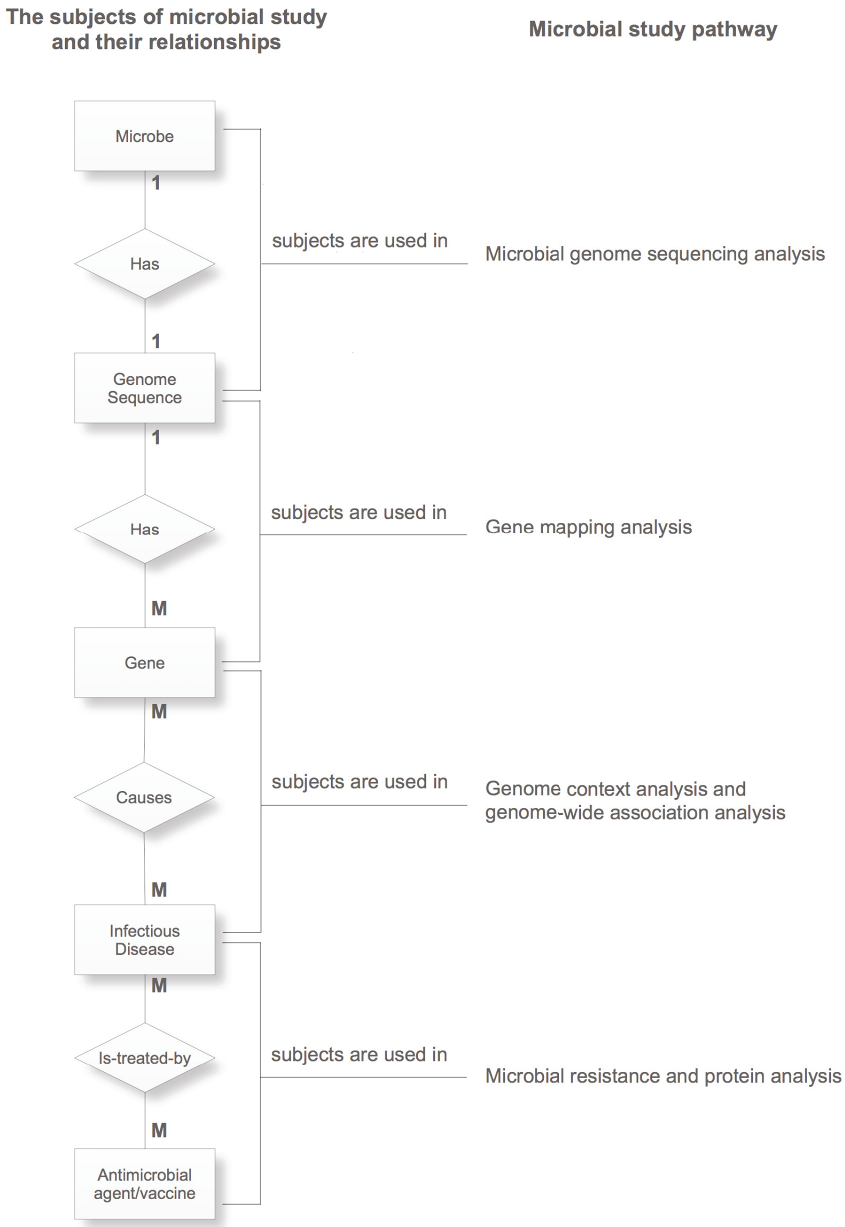

The subjects of microbial study and their relationships
Microbial study pathway
Microbe
1
Has
1
Genome Sequence
1
Has
M
Gene
M
Causes
M
Infectious Disease
M
Is-treated-by
M
Antimicrobial agent/vaccine
subjects are used in
Microbial genome sequencing analysis
subjects are used in
Gene mapping analysis
subjects are used in
Genome context analysis and genome-wide association analysis
subjects are used in
Microbial resistance and protein analysis

Fig. 1. The Relationship Between the Various Subjects of the Microbial Study and the Analyses in the Microbial Study Pathway

4.1 Answering PBC Measurement Request in Microbial Genomics

In handling PBC measurement requests for the microbial genomics domain, PBC measurement providers need to configure the elements of the PBC model that are specific for this domain. To describe the configuration needed for reference populations, assume that the microbial universe consists of reference populations whose individuals are from databases (CSs) in the microbial domain identified by domain experts. PBC measurement providers need to configure the form of reference population table schema, together with the information that must be stored within the tables. The basic configuration of the PBC model defines the type of reference population table schema (called the POPSCHEMA) to be in the general form of: $\langle I, source, A\rangle$ (where I is identifier attribute(s), *source* is the name of the source attribute and A is the set of attributes other than I and *source*) as as a form of schema to support all types of PBC measurement requests.

Suppose that the general form of the reference population table schema is adopted and the reference candidate gene population is configured as a table called `gene` with schema: $\langle\texttt{geneId}, \texttt{source}, \texttt{species}\rangle$. In addition to the reference population, we also need to configure the microbial universe, its CSs and the population maps that are specific for this domain. Based on the basic configuration of the PBC model, two variables can be defined namely UP (the set of reference population tables in the universe and their schema) and CS (the set of CSs in the universe). The following is an example of the instances of the variables configured and stored for PBC measurement in the microbial domain:

- UP=$\{$`(gene,`$\langle$`geneId, source, species`$\rangle$`)`$\}$,
- CS=$\{$`(CMR, http://www.tigr.org/CMR,` PM_{CMR} `)`$\}$, where PM_{CMR} is a set of population maps for Comprehensive Microbial Resource (CMR) in the form of:
  $\{$`gene, SELECT geneCode, speciesCode FROM microbeGene`$\}$. Every instance of PM_{CMR} consists of the name of the population table (that is equivalent to the name of the reference population it contributes), and the query against the table in the CS.
  For brevity, we only show an instance for each variable.

Assuming that all elements of PBC have been configured for the microbial domain, we will next present one type of PBC measurement requests that can be supported by the PBC configuration. For a request to measure completeness of a candidate gene population relative to the reference candidate gene population that consists of genes coming from certain CSs of the microbial universe only (e.g., CMR and SwissProt), PBC is measured as:

$$Completeness(\,\langle\texttt{ExtGENE}\rangle\,,\langle\texttt{gene}, COND\rangle) = \frac{|\texttt{ExtGENE}\cap(\Pi_{\texttt{key(gene)}}(\sigma_{COND}\texttt{gene}))|}{|\Pi_{\texttt{key(gene)}}(\sigma_{COND}\texttt{gene})|},$$

where `ExtGENE` is the external gene data set under measure, `key(gene)` is a function that retrieves `geneId` (the identifier of genes) from `gene`, *COND* is a conjunction of conditions on `gene` using *source* attribute as the predicate. For example, one condition in *COND* is specified as `source IN ('CMR','SwissProt')` in the query.

This type of request could be driven by the need to use a reference gene population that comes from a preferred source e.g. based on trust/reputation. Because not all CSs chosen by the PBC measurement provider are preferred by the person requesting the measurement, we need to filter the genes by specifying the condition on the `source` predicate in the query. Other specific queries can be specified by adding the required predicate(s) in *COND* (e.g., the analysis may interested in genes for certain microbe species called *S.bongori*).

5 Conclusion

In conclusion, we discovered that defining what is the 'complete' reference data set (the population) to use can be difficult such as in the microbial genomics case. In this paper, the elements of PBC have been defined, and we found that the choice of using true populations (which are the true, complete reference data sets) is often hindered by the lack of knowledge of the true population individuals and technical issues (i.e. accessibility of data sources). Using approximate populations however is complicated by the task of gathering population individuals from multiple data sources that would contribute to the closest approximation of the true populations. How practical is the PBC model is one of the remaining questions that call for further investigation in our future work.

Acknowledgments. The authors would like to thank the financial assistance provided by the Universiti Teknikal Malaysia, Melaka (UTeM) and the Ministry of Higher Education, Malaysia during the course of this research and the members of Information Management Group (IMG), The University of Manchester for their constructive comments.

References

1. Pipino, L.L., Lee, Y.W., Wang, R.Y.: Data quality assessment. Communications of the ACM 45, 211–218 (2002)
2. Iles, M.M.: What can genome-wide association studies tell us about the genetics of common disease. PLOS Genetics 4, 1–8 (2008)
3. Tiffin, N., Andrade-Navarro, M.A., Perez-Iratxeta, C.: Linking genes to diseases: it's all in the data. Genome Medicine 1, 1–7 (2009)
4. Codd, E.F.: Extending the database relational model to capture more meaning. ACM Transactions on Database Systems (TODS) 4 (1979)
5. Reich, D.E., Gabriel, S., Atshuler, D.: Quality and completeness of SNP databases. Nature Genetics 33, 457–458 (2003)
6. Zaniolo, C.: Database relations with null values. Journal of Computer and System Sciences 28, 142–166 (1984)
7. Codd, E.F.: Understanding relations (installment #7). Bulletin of ACM SIGMOD 7, 23–28 (1975)
8. Imieliński, T., Lipski, J.: Incomplete information in relational databases. Journal of the ACM 31, 761–791 (1984)

9. Fox, C., Levitin, A., Redman, T.: The notion of data and its quality dimensions. Information Processing and Management 30, 9–19 (1994)
10. Motro, A.: Integrity = validity + completeness. ACM Transactions on Database Systems 14, 480–502 (1989)
11. Motro, A., Rakov, I.: Estimating the Quality of Databases. In: Andreasen, T., Christiansen, H., Larsen, H.L. (eds.) FQAS 1998. LNCS (LNAI), vol. 1495, pp. 298–307. Springer, Heidelberg (1998)
12. Sampaio, S.F.M., Sampaio, P.R.F.: Incorporating completeness quality support in internet query systems. In: CAiSE Forum. CEUR-WS.org, pp. 17–20 (2007)
13. Scannapieco, M., Batini, C.: Completeness in the relational model: a comprehensive framework. In: Ninth International Conference on Information Quality (IQ), pp. 333–345. MIT (2004)
14. Knudson, A.: Mutation and cancer: statistical study of retinoblastoma. Proceedings of the National Academy of Sciences of the United States of America 68, 820–823 (1971)
15. Hashimoto, C.: Population census of the chimpanzees in the Kalinzu forest, Uganda: Comparison between methods with nest counts. Primates 36, 477–488 (2006)
16. Liang, Z., Ma, Z.: China's floating population: new evidence from the 2000 census. Population and Development Review 30, 467–488 (2004)
17. Bird, A., Tobin, E.: Natural kinds. In: The Stanford Encyclopedia of Philosophy (summer 2010)
18. Science Daily: Human gene count tumbles again (2008), `http://www.sciencedaily.com/releases/2008/01/080113161406.htm` (accessed June 27, 2011)
19. Maddux, R.: The origin of relation algebras in the development and axiomatization of the calculus of relations. Studia Logica 50, 421–455 (1991)
20. Falkow, S.: Who speaks for the microbes? Emerging Infectious Disease 4, 495–497 (1998)
21. Fraser, C.M., Eisen, J.A., Salzberg, S.L.: Consanguinity and susceptibility to infectious diseases in humans. Nature 406, 799–803 (2000)

Road Traffic Prediction Using Context-Aware Random Forest Based on Volatility Nature of Traffic Flows

Narjes Zarei, Mohammad Ali Ghayour, and Sattar Hashemi

Department of Computer Science and Engineering, Shiraz University, Shiraz, Iran
{n_zarei,a_ghayour,s_hashemi}@shirazu.ac.ir

Abstract. Nowadays short-term traffic prediction is of great interest in Intelligent Transportation Systems (ITS). To come up with an effective prediction model, it is essential to consider the time-dependent volatility nature of traffic data. Inspired by this understanding, this paper explores the underlying trend of traffic flow to differentiate between peak and non-peak traffic periods, and finally makes use of this notion to train separate prediction model for each period effectively. It is worth mentioning that even if time associated with the traffic data is not given explicitly, the proposed approach will strive to identify different trends by exploring distribution of data. Once the data corresponding trends are determined, Random Forest as prediction model is well aware of data context, and hence, it has less chance of getting stuck in local optima. To show the effectiveness of our approach, several experiments are conducted on the data provided in the first task of 2010 IEEE International Competition on Data Mining (ICDM). Experimental results are promising due to the scalability of the proposed method compared to the results given by the top teams of the competition.

Keywords: Intelligent transportation systems (ITS), urban traffic congestion, short-term prediction, random forest.

1 Introduction

Traffic congestion is a worldwide concern that influences different aspect of urban life. Regard to the land and cost limitation, expanding existing infrastructure is not always feasible. Therefore, Intelligent Transportation Systems (ITS) was developed to improve the transportation system by smarter use of capacity of existing transport infrastructures. To augment urban road management, ITS technologies rely not only on real-time traffic data, but also on future predictions of traffic conditions. Accordingly, accurate traffic prediction techniques are essential to estimate future traffic conditions, usually from several minutes to hours ahead. As a result, Advanced Traveler Information Systems (ATIS) and Advanced Traffic Management Systems (ATMS) would be able to apply traffic control strategies to prevent traffic congestion and allow travelers to optimize their routes which result in an efficient, safe and sustainable transport system with less road congestion. Among a great variety of methodologies developed for short-term traffic predictions such as Kalman filtering [1, 2], nonparametric statistical methods [3, 4], sequential learning [5] and Neural Network models [6, 7], Times Series analyses [8, 9] are the most commonly used approaches in

A. Selamat et al. (Eds.): ACIIDS 2013, Part I, LNAI 7802, pp. 196–205, 2013.

short-term traffic prediction. Lee and Fambro [10] developed the subset autoregressive integrated moving average (ARIMA) model and compared it with full autoregressive (FAR), subset autoregressive (SAR) and full ARIMA models. Results showed that among other time-series models, a subset ARIMA model delivered the most accurate and stable results. In 2005, Generalized Auto-Regressive Conditional Heteroscedasticity (GARCH) model was used for representing the dynamics of traffic flows volatility [11]. Indeed, their purpose was to provide better confidence intervals for predictions, with regards to the variance changes of traffic flow through time. Even though traffic data is considered as time series data, due to time-dependent nature of the data, applying classical Time Series analyses needs carful preprocessing. In 2003, Washington et al. [12] suggested to remove time variation of traffic volume in order to achieve a "flatter" power spectrum. However, this approach may fails due to loosing valuable information useful for short-term prediction.

Recently, tendency of recent researches have been oriented toward data mining algorithms which are cable of mining knowledge from huge databases, like traffic data. Among them, classification and regression trees are widely used with considerable efficiency [13, 14]. In 2010, Benjamin Hamner [13] conducted a study using an ensemble of Random Forests. Different downsampling and resampling techniques were also performed to predict traffic volume of road segments for 10-20 minutes ahead. The drawback of this method is overlooking the underlying traffic pattern. Indeed, most studies concentrated only on applying different learning algorithms on historical traffic flows. However, with regard to the time-dependent volatility nature of traffic flows, it is essential to explore the underlying traffic flow before building traffic prediction models. Cluster analysis models which capture the trend of traffic flow have been commonly used with the purpose of effective modeling [15, 16]. However, in this study, it is shown that extracting just peak periods traffic flows result in more accurate outcomes.

In the current work, since the time associated with the traffic peak is not available, the proposed method's aim is to first identify the time of the traffic flows by examining the distribution of data. Separate context-aware Random Forests are then trained for peak and non-peak period context.

The proposed method was evaluated using a time series of simulated congestion measurements from ten selected road segments of Warsaw. The data is taken from the first task of 2010 IEEE International Competition on Data Mining (ICDM) [17]. The main purpose is to predict future traffic congestion based on historical ones. Experimental results indicate promising improvements comparing to the models conducted over the same dataset.

The remainder of this paper is organized as follows: section 2 introduces problem setting and data description .Proposed method is explained in section 3, while section 4 presents evaluation of the method using the traffic congestion data set and the achieved results. Finally, conclusion is presented in section 5.

2 Data Description

Traffic data collection commonly relies on a network of sensors which provide standard traffic parameters such as traffic volume, travel speed and occupancy. Currently, the dominant technology employed for traffic data collection is Automatic Traffic Recorders (ATR). ATRs are magnetic loops embedded underground, capable of counting vehicles as they pass through. The data gathered through ATRs is used further to make short-term and long-term predictions, for the purpose of driver navigation and urban planning which was especially addressed in first task of 2010 IEEE ICDM Traffic Prediction competition. The dataset was acquired from a highly realistic simulator of vehicular traffic, Traffic Simulation Framework (TSF) - a complex tool for simulating and investigating vehicular traffic in cities- developed at the University of Warsaw [18]. The simulations used authentic map of Warsaw, Poland, and it is expected in the future that TSF would be able to work with maps of any city. In this simulator, different rules were considered to model crossroads, traffic lights and multi-lane streets to reproduce accurate vehicular traffic phenomenon.

To provide the training dataset, the simulation was run for 1000 hours, divided into a hundred of 10-hour long independent cycles. Furthermore, the data is assembled into one-minute interval and form 6000 records. Each record contains congestion number of cars for two directions of ten selected road segments of Warsaw, meaning each record has 20 values. Each record can be represented as bellow:

$$V_t = \ < V_{t,1}, V_{t,2}, \cdots, V_{t,r}> \qquad 1 \leq t \leq 60000\ , \quad r = 20 \tag{1}$$

Where V_t represent t^{th} record and $V_{t,r}$ refers to the traffic volume of r^{th} road during t^{th} minute. Another set of data including 1000 hours of traffic simulation divided into 60-minute long windows was used as test set, of which only the first 30 minutes are used as test set, while the other 30 minutes were left for evaluation of predictions. Given the first half hour of each test window, the objective is to predict total of number of cars for time periods between 10 and 20 minutes ahead, i.e., for the period between 41'st and 50th minute of that window as the target. The test window and corresponding target vector are shown as below:

$$\text{Test Window} = < V_{t,1}, V_{t,2}, \ldots, V_{t,20}> \qquad 1 \leq t \leq 30 \tag{2}$$

$$\text{Target Vector} = \quad < \textstyle\sum_{t=41}^{50} V_{t,1}\ , \sum_{t=41}^{50} V_{t,2}\ , \ldots\ldots\ , \sum_{t=41}^{50} V_{t,20}\ > \tag{3}$$

It is worth to note that, windows in test dataset were permuted, so that it was not possible to deduce future congestion by looking at the following time window. To get more elaborated view into the terms which are used along the body the article, refer to Table 1.

Table 1. Elaborated view into the terms used along the body of manuscript

Training data	100 cycles
Each Cycle	10 hours
Each Hour	60 records (minutes)
Test data	1000 windows
Each window	30 minutes

3 Proposed Method

To come up with more precise traffic prediction, the proposed method is consisted of three steps: (1) first a preprocessing step was applied on the data to decrease its dimension with less duplicate information and then (2) with regards to the observed traffic flow trends of the data, the dataset was differentiated into peak and non-peak period context and finally (3) two separate context-aware Random forests are then trained for predicting traffic congestion.

Traffic flow of urban roads usually follows a time-dependent trend pattern. The observed trend for a given road is usually the same during different days. In addition, a common feature of all roads is existence of high traffic volume during rush hours.

The observed trend during ten typical 10-hour long cycles of given dataset is shown in Fig. 1 which is achieved by a simple averaging between all road segments.

As it is shown, the sharp peaks occur during the first hours of all cycles. Although the time associated with the trends is not given in this dataset, but with regards to the explanation given on the data simulation [18], these independent cycles can be considered as different days and the first hours of cycles reflex the rush hours of days - (morning or evening peak periods). To be specific, it was stated that after each cycle, distribution of source and destination of vehicles were randomly selected and simulation started from the scratch. Therefore, first hours of cycles can be interpreted as the concentration of trips such as commuting to work and school from homes in the morning. In sum, observing different trends within the traffic flows lead us to conclude that, no particular analytical model should be trained on all the data exclusively. In order to exploit the achieved information in building the model, the steps were considered as shown in Fig. 2.

Initially, in order to fit a model, there is a need to extract a feature matrix X and target matrix Y from the training data file. Considering each 60-minute long window in the training file, the first half is used to extract features and the summation of

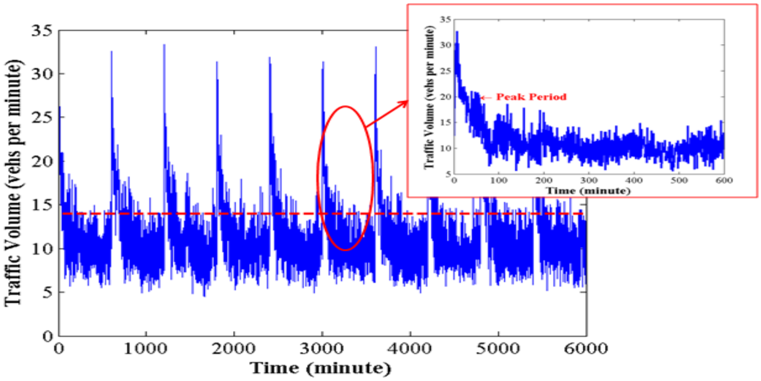


Fig. 1. Observed trend of traffic flows of ten typical 10-hour long cycles. Peak periods are identified during first hours of cycles which are magnified. The trend is obtained by a simple averaging between all road segments.

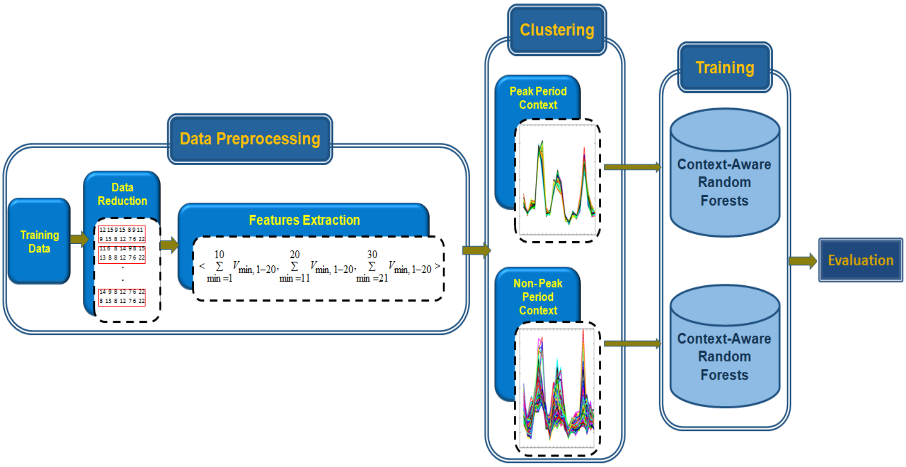


Fig. 2. The processes involved in the proposed methodology which are followed before training context-aware random forest with the purpose of short-term prediction

41th -50th minutes of that window used as the prediction target vectors. Due to the minute level presentation of data, a level of aggregation is applied to aggregate data into longer minutes-blocks to avoid high level of fluctuations and dimensionality. The size of aggregation level should be long enough to overlook duplicated information. Experimental result shows that a time window of 10-minutes aggregation yields the best performance. As a result, 3 vectors of 20-values were delivered per window, which formed a feature vector consisting of 60 values. Feature and target vectors are represented in the following format:

$$\text{Features Vector} =< \sum_{\min=1}^{10} V_{\min,1-20}, \sum_{\min=11}^{20} V_{\min,1-20}, \sum_{\min=21}^{30} V_{\min,1-20} > \quad (4)$$

$$\text{Target Vector} =< \sum_{\min=41}^{50} V_{\min,1-20} > \quad (5)$$

Clearly, extracting one feature from each one hour-long window, deliver a feature matrix of 1000 training data points, 10 for each of 100 10-hour training cycles. Y is a 1000-by-20 matrix that each column corresponding to the summation of congestion on each road segments.

After the preprocessing step, cluster analysis should be applied to traffic flow to differentiate between peak and non-peak traffic periods. Although a full cluster analysis is commonly performed in the studies, in this study it was observed that discriminating just peak and non-peak period contexts result in more accurate outputs. Since the time associated with the traffic trend is not given in this study, a proper similarity measure is needed to determine these observations of different contexts. As it was investigated, the range of traffic volumes related to the different contexts is significantly different. Hence, the applied metric should be shift variant to be able to

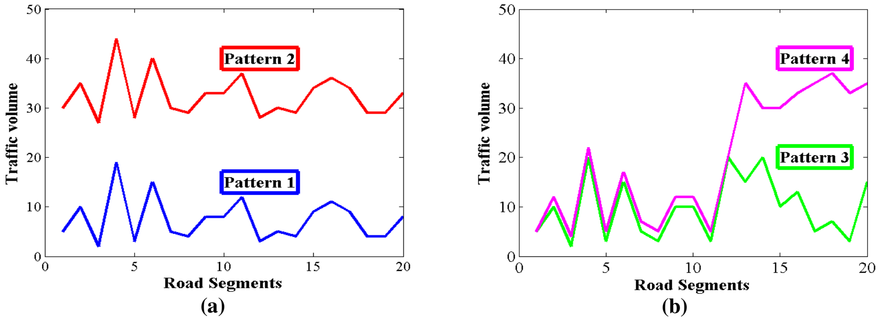


Fig. 3. Illustration to indicate requirements of a suitable similarity metric. It should be shift variant (a) and also should not consider partial matching (b) to put all these patterns in different context.

discriminate them. To be more illustrative, in Fig. 3(a), patterns 1 and 2 deem similar, however, the metric should not put them in the same context. On the other hand, the applied metric should not do partial matching. The main reason is that each segment has a specified range of traffic volume which increased by specified factor even during peak periods. For more illustration, Fig.3 (b) shows two patterns as an example which should not be placed in the same context.

Results prove that examining other choice of distance metrics deliver no significant improvements more than Euclidean metric which calculate the distance between two observations X and Y as follow:

$$Euclidean(X,Y) = \sqrt{\sum_i (X_i - Y_i)^2} \tag{6}$$

As a result of applying the metric, peak period and non-peak period contexts are delivered and visualized in Fig. 4(a) and (b), respectively. As it is clear, majority of observations of peak period context behave very similarly. To be more specific, consider road segment 3 in both context; the range of traffic volume recorded for this segment is about (19-24) in different observations of peak periods context. However, in non-peak context, this segment experiences different traffic volumes which don’t fall into the range as limited as in peak period observations. Therefore, with regard to the specific trend of road segments during peak period time, it is suggested to discriminate them from other observations. Once the data corresponding trends are determined, Random Forest as prediction model is well aware of data context, and hence, training would be done separately on more similar traffic flows and the effect of uncorrelated observations on prediction inaccuracy is prevented.

Supervised classification data mining algorithms: -Random forest (Leo Brieman, 2001) [19] is a general variant of bagging methods using a collection of CART-like trees as a base classifier. Final prediction computed by averaging the output (for regression) or voting (for classification). The main remarkable features of Random Forest are its efficiency on large datasets. As such, it is strong in both regression and classification.

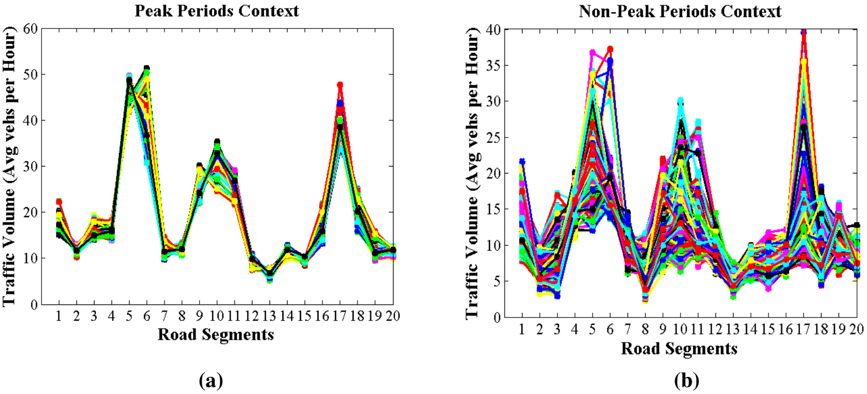


Fig. 4. Observed traffic flows of all 20 road segments divided into peak periods (a) and non-peak period (b) contexts. Majority of observations of peak context behave very similarly. Different colors refer to different observations.

4 Experimental Results

As described earlier, given the traffic congestion of first 30-minute of 1000 test windows, the goal is to predict congestion for the period 41th – 50th of next half-hour of next half-hour of that window. To evaluate the prediction accuracy, root-mean-square error (RMSE) was used to measure the differences as follows:

$$\mathrm{RMSE} = \sqrt{\frac{\sum_{i=1}^{N}(\mathrm{y_i}-\hat{\mathrm{y}}_\mathrm{i})}{N}} \tag{7}$$

Where, $\mathrm{y_i}$ is the observed traffic volume; $\hat{\mathrm{y}}_\mathrm{i}$: is the predicted traffic volume and N is the number of observation points. Matlab R2012a was used to perform all the analysis in this study including the Random Forest modeling as well as distribution analysis Random forest was trained using 70 trees, minimum leaf size 5 and number of variables for random feature selection was set to one third of the total number of variables.

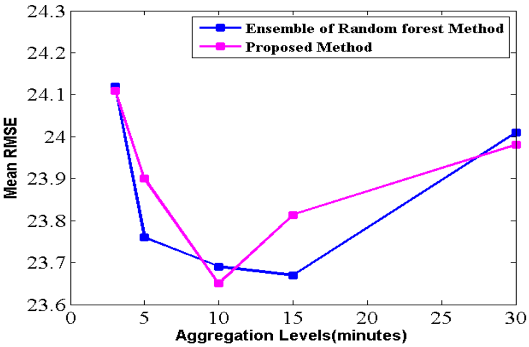


Fig. 5. The mean RMSE of proposed method compared with Ensemble of Random Forests approach. The aggregation level was varied from 3 to 30 minutes.

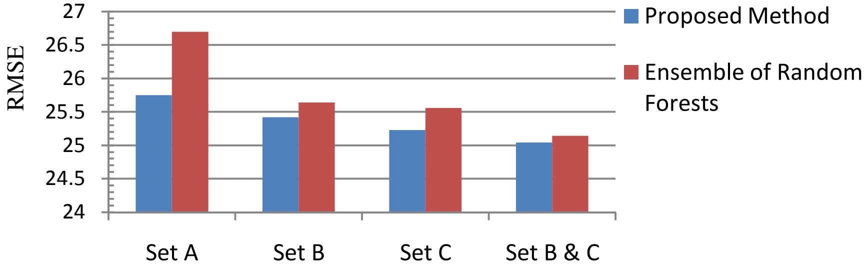


Fig. 6. RMSE comparison of proposed method and Ensemble of Random Forest model, applied on sets A, B and C consisted of 1000, 11000 and 55000 training points respectively. The proposed method dramatically reduces the error even on the most scalable set (A).

As mentioned earlier, there is a need to apply an aggregation phase over the given data. Since in the test sets, we are faced to 30-minute long windows, the possible aggregation levels are 3, 5, 6, 10, 15 and 30 which led to feature vectors of 200, 120, 100, 60, 40 and 20 features, respectively. Fig .5 demonstrates the manner in which the mean RMSE changes with regards to different level of aggregation where a time window of 10-minutes aggregation yields the best performance.

Considering the large scale of given traffic data, scalability of developed methods is still problematic. In [13], different sets of feature matrix with different size were extracted from the training data. Set A, B and C cover 1000, 11000 and 55000 training points, respectively, while, the final model was made based on an ensemble of set B and C. As it was also stated in [13], the primary limitation of the method algorithm was that, while it worked well for 20 ATRs, it did not scale well to thousands of ATRs as it is implemented.

In Fig. 6, the utility of proposed method conducted over the test dataset is compared with what was achieved by our implementation of Ensemble Random Forest method [13] which was developed by the third winner of the first track of ICDM 2010 contest. As it is shown, the developed methodology is able reduces the error dramatically even on the set A, which is more scalable. However, other sets are applied to improve the performance. Our experimental results show remarkable improvement of accuracy to support the idea.

5 Conclusion

It is widely accepted that ITS technology has been brought efficient solution in transportation system. In this regard, accurate and timely short-term traffic flow predictions are an essential component to support traffic management centers. Therefore, in the last two decades, much research efforts have been concentrated on developing effective prediction models. However, with regard to the time-dependent nature of traffic data, the underlying data should be explored prior to building models.

In this article, a methodology framework differentiates between traffic flows related to peak and non-peak periods. Commonly, the peak period's traffic flow are identified regarding to their occurrence time, however, herein, we are not given the associated time. Therefore, the whole distribution of data was examined to detect different contexts. So, our scheme comes up to train context-aware Random Forest for more accurate predictions. Empirical results prove the method's effectiveness and scalability.

References

1. Okutani, I., Stephanedes, Y.J.: Dynamic prediction of traffic volume through Kalman filtering theory. Transportation Research Part B 18(1), 1–11 (1984)
2. Whittaker, J., Garside, S., Lindveld, K.: Tracking and predicting a network traffic process. International Journal of Forecasting 13, 51–61 (1997)
3. Davis, G.A., Nihan, N.L.: Nonparametric regression and short-term freeway traffic forecasting. ASCE Journal of Transportation Engineering 117(2), 178–188 (1991)
4. Smith, B.L., Williams, B.M., Oswald, R.K.: Parametric and nonparametric traffic volume forecasting. Presented at the 2000 Transportation Research Board Annual Meeting, Washington, DC (2000)
5. Chen, H., Grant-Muller, S.: Use of sequential learning for short-term traffic flow forecasting. Transportation Research Part C 9, 319–336 (2001)
6. Chang, S.C., Kim, S.J., Ahn, B.H.: Traffic-flow forecasting using time series analysis and artificial neural network: the application of judgmental adjustment. Presented in the 3rd IEEE International Conference on Intelligent Transportation Systems (2000)
7. Park, D., Rilett, L.R.: Forecasting multiple-period freeway link travel times using modular neural networks. Transportation Research Record 1617, 63–70 (1998)
8. Ghosh, B., Basu, B., O'Mahony, M.: Multivariate short-term traffic flow forecasting using time-series analysis. IEEE Transactions on Intelligent Transportation Systems 10(2), 246–254 (2009)
9. Nihan, N.L., Holmesland, K.O.: Use of the box and Jenkins time series technique in traffic forecasting. Transportation 9 (2) (1980)
10. Lee, S., Fambro, D.B.: Application of Subset Autoregressive Integrated Moving Average Model for Short-Term Freeway Traffic Volume Forecasting. Transportation Research Record, No.1678, pp.179–188 (1999)
11. Kamarianakis, Y., Kanas, A., Prastacos, P.: Modeling Traffic Volatility Dynamics in an Urban Network. Transportation Research Record: Journal of the Transportation Research Board No. 1923, 18–27, Transportation Research Board (2005)
12. Washington, S.P., Karlaftis, M.G., Mannering, F.L.: Statistical and Econometric Methods for Transportation Data Analysis. Chapman and Hall/CRC Press, Boca Raton (2003)
13. Hamner, B.: Predicting Future Traffic Congestion from Automated Traffic Recorder Readings with an Ensemble of Random Forests. In: 2010 IEEE International Conference on Data Mining Workshops (ICDMW), pp. 1360–1362. IEEE (December 2010)
14. Gil Bellosta, C.J.: A convex combination of models for predicting road traffic. In: 2010 IEEE International Conference on Data Mining Workshops (ICDMW). IEEE (2010)

15. Qi, Y.: Probabilistic models for short term traffic conditions prediction. Submitted to the Graduate Faculty of the Louisiana State University and Agricultural and Mechanical College in partial fulfillment of the requirements for degree of Doctor of Philosophy (May 2010)
16. Vlahogianni, E.I.: Enhancing Predictions in Signalized Arterials with Information on Short-Term Traffic Flow Dynamics. Journal of Intelligent Transportation Systems 13(2), 73–84 (2009)
17. Wojnarski, M., Gora, P., Szczuka, M., Hung, N.S., Swietlicka, J., Zeinalipour, D.: IEEE ICDM 2010 Contest: TomTom Traffic Prediction for Intelligent GPS Navigation. In: 2010 IEEE International Conference on Data Mining Workshops (ICDMW), pp. 1372–1376. IEEE (December 2010)
18. Gora, P.: Traffic Simulation Framework. In: 2012 UKSim 14th International Conference on Computer Modelling and Simulation (UKSim), pp. 345–349. IEEE (March 2012)
19. Breiman, L.: Random Forests. Machine Learning 45(1), 5–32 (2001)

Scoring-Thresholding Pattern Based Text Classifier

Moch Arif Bijaksana, Yuefeng Li, and Abdulmohsen Algarni

School of Electrical Engineering and Computer Science,
Queensland University of Technology, Brisbane, Australia
moch.bijaksana@student.qut.edu.au,
{y2.li,a1.algarni}@qut.edu.au

Abstract. A big challenge for classification on text is the noisy of text data. It makes classification quality low. Many classification process can be divided into two sequential steps scoring and threshold setting (thresholding). Therefore to deal with noisy data problem, it is important to describe positive feature effectively scoring and to set a suitable threshold. Most existing text classifiers do not concentrate on these two jobs. In this paper, we propose a novel text classifier with pattern-based scoring that describe positive feature effectively, followed by threshold setting. The thresholding is based on score of training set, make it is simple to implement in other scoring methods. Experiment shows that our pattern-based classifier is promising.

Keywords: Text classification, Pattern mining, Scoring, Thresholding.

1 Introduction

An important problem for text classification is the classification quality (represented with F_1 measure) is low, because of noisy text corpus. This problem becomes more severe in corpus with few positive and small number of training set. Important issues related to this problem are how to describe positive feature and how to set a suitable threshold.

SVM, Rocchio and kNN are existing popular effective text classification methods [9]. In these classifiers, classification process can be divided into two sequential steps scoring and threshold setting (thresholding).

In real life, many classification problems are multi-class and multi-label. Multi-class and multi-label classification is popularly solved the problem by splitting into several binary classifications. SVM and Rocchio usually apply this approach. Binary classification theoretically is more generic than multi-class classification or multi-label classification [9]. In this paper we use binary dataset for experiments.

Typically, scoring process is conducted by the classifier and thresholding is a post-processing. For classification, thresholding is often considered as a trivial process and is not important; therefore it is under-investigated. However, Yiming Yang [12] proved that thresholding is important and not simple. She proved that an effective thresholding strategy produces significantly better classfication effectiveness significantly than other thresholding strategies.

We propose a new two-stage model which set both score and threshold values explicitly. In scoring stage, we focus on describing positive feature by using patterns,

A. Selamat et al. (Eds.): ACIIDS 2013, Part I, LNAI 7802, pp. 206–215, 2013.

statistical semantic features that captures semantic relations between terms. While in thresholding stage we use an effective training-based model.

We conducted substantial experiments on popular text classification corpus based on Reuters Corpus Volume 1 (RCV1), compared with SVM and Rocchio to evaluate the proposed model. The results show that our model is promising.

The rest of this paper is structured as follows: Section 2 discusses related work. Section 3 and 4 proposes the techniques of scoring and thresholding respectively. The experiment design is described in Section 5, whereas the results are discussed in Section 6. Finally, Section 7 gives concluding remarks.

2 Related Work

Classification can be done in two ways, new documents is directly predicted the class, or scored (or ranked) [9]. Both are generally performed in two stages scoring and thresholding. Scoring process conducted by the classifier and thresholding is a post-processing. Information retrieval models are the basis of ranking algorithm that is used in search engines to produce the ranked list of documents [2].

Scoring for ranking is the main problem in information filtering field, the objective is to effectively score incoming documents to rank. Some recent work in information filtering are including [4].

Different to the ranking task, classification is the task of assigning documents to predefined categories. A comprehensive review of text classification methods can be found in [9]. To date, many classification methods, such as Naive Bayes, Rocchio, kNN and SVM have been developed in IR [6].

For classification, thresholding is often considered a trivial process and is not important; therefore it is under-investigated. However, Yiming Yang [12] proved that thresholding is important and not simple. She proved, using kNN, an effective thresholding strategy produces significantly better classfication effectiveness significantly than other thresholding strategies.

Existing works on thresholding strategy are generally in the context of post-procesing classification of or for multi-label classification problems such as [3,12]. However, in principle, these thresholding strategies can be used for tranforming ranking to binary decision classification. To our knowledge, only a few number of works focused on thresholding strategies for ranking into binary decision, among others [5,13].

There are two steps in rank to binary decision transformation, firstly scoring documents $score(d_j, c_i)$, then thresholding [9]. Most of classification are processed in two steps scoring and thresholding as well. These classifier usually use default threshold, for example threshold score=0 for SVM and probability = 0.5 for Bayes classifier [3]

There are at least three popular thresholding strategies, namely ranking-based, score-based, and proportional-based [9,12,8].

3 Pattern-Based Scoring Model

In this paper, we assume that all documents are split into paragraphs. So a given document d_i yields a set of paragraphs $PS(d_i)$. In our model, we use sequential closed patterns. The definition of sequential closed pattern can be found in [4].

Table 1. Pattern based document representation

Doc	Patterns
d_1	$\langle carbon\rangle_4, \langle carbon, emiss\rangle_2, \langle air, pollut\rangle_2$
d_2	$\langle greenhous, global\rangle_3, \langle emiss, global\rangle_2$
d_3	$\langle greenhous\rangle_2, \langle global, emiss\rangle_2$
d_4	$\langle carbon\rangle_3, \langle air\rangle_3, \langle air, antarct\rangle_2$
d_5	$\langle emiss, global, pollut\rangle_2$

Table 1 illustrates document representation in our pattern-based model, which is based on Relevance Feature Discovery (RFD) [4] and Pattern Taxonomy Model (PTM) [11]. In this figure d_1 has three pattern features $\langle carbon\rangle_4$, $\langle carbon,\ emiss\rangle_3$, and $\langle air, pollut\rangle_2$. Subscripted values are support values which represents weight. It means that in d_1 there are four paragraphs contain pattern $\langle carbon\rangle$, three paragraphs contain pattern $\langle carbon, emiss\rangle$, and two paragraphs contain pattern $\langle air, pollut\rangle$. A termset X is called a frequent sequential pattern if its relative support $supp_r(X)$ is greater than or equal to a predefined minimum support, that is, $supp_r(X) \geq min\_sup$.

RFD is applied as effective information filtering system. Different to most other document representation, instead of as input for machine learning classifier, RFD representation is then used to produce a set of weighted term. The set of weighted term is used as class representation, it makes scoring process (for new documents) more efficient.

In pattern-based, the class representation is in the form of a set of weighted terms. The number of terms in class representation is relatively small compared to the size of the vocabulary in the class.

The terms weight in class representation is calculated from the appearance of terms in the document representation (patterns). There are several ways to calculate the weight of term [4,11]. The basic weight of term t in dataset D^+_{tr} is

$$weight(t, D^+_{tr}) = \sum_{i=1}^{|D^+_{tr}|} \frac{|\{p | p \in SP_i, t \in p\}|}{\sum_{p \in SP_i} |p|}$$

where SP_i is pattern set of pattern p in document d_i, and $|p|$ is the number of term in pattern p. For example in Table 1, $D^+_{tr} = \{d_1, d_2, \ldots, d_5\}$, term *global* (which appears in document d_2, d_3, $\ldots d_5$), has $weight(global, D^+_{tr}) = \frac{2}{4} + \frac{1}{3} + \frac{1}{3} = \frac{7}{6}$.

However, for scoring new documents, different to Rocchio (which uses similarity or distance measure to score new documents), pattern-based simply sum the term weights appear in new documents [4].

Then we group terms into three groups (positive specific terms, general terms and negative specific terms) based on their appearances in a training set. Given a term $t \in T$, its $coverage^+$ is the set of positive documents that contain t, and its $coverage^-$ is the set of negative documents that contain t. We assume that terms frequently used in both positive documents and negative documents are general terms. Therefore, we want to classify terms that are more frequently used in the positive documents into the positive specific category; and the terms that are more frequently used in the negative documents into the negative specific category.

Based on the above analysis, we define the *specificity* of a given term t in the training set $D = D^+ \cup D^-$ as follows:

$$spe(t) = \frac{|coverage^+(t)| - |coverage^-(t)|}{n}$$

where $coverage^+(t) = \{d \in D^+ | t \in d\}$, $coverage^-(t) = \{d \in D^- | t \in d\}$, and $n = |D^+|$. $spe(t) > 0$ means that term t is used more frequently in positive documents than in negative documents.

We present the following classification rules for determining the general terms G, the positive specific terms T^+, and the negative specific terms T^-:

$$G = \{t \in T | \theta_1 \leq spe(t) \leq \theta_2\},$$

$$T^+ = \{t \in T | spe(t) > \theta_2\},\ and$$

$$T^- = \{t \in T | spe(t) < \theta_1\}.$$

where θ_2 is an experimental coefficient, the maximum bound of the specificity for the general terms, and θ_1 is also an experimental coefficient, the minimum bound of the specificity for the general terms. We assume that $\theta_2 > 0$ and $\theta_2 \geq \theta_1$.

The initial weights of terms are revised according to the following principles: increment the weights of the positive specific terms, decline the weights of the negative specific terms, and do not update the weights of the general terms. The details are described as follows:

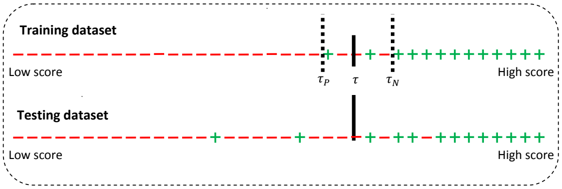


Fig. 1. Threshold setting.

$$weight(t) = \begin{cases} w(t) + w(t) \times spe(t), & \text{if } t \in T^+ \\ w(t), & \text{if } t \in G \\ w(t) - |w(t) \times spe(t)|, & \text{if } t \in T^- \end{cases}$$

where w is the initial weight.

Score value of a document d_i is

$$score(d_i) = \sum_{t_j \in T} weight(t_j)$$

4 Thresholding Model

The threshold value (τ) in our model is based on score of document. Score in document d_i, $Score(d_i)$, is its weight in RFD model. For thresholding, score can be based on training set, validation set, or testing set. For a dataset with a small number of training set, it is difficult to get a representative validation set. While using testing set for thresholding, make thresholding model is not suitable for online learning. Our model generate threshold based on score of training set D_{tr}, which consists of a set of positive documents, D_{tr}^{+}, and a set of negative documents, D_{tr}^{-}.

Our τ is based on minimum score of positive training document (τ_P) and minimum score of negative training document (τ_N) (see Figure 1)

$$\tau_P = \min_{d_i \in D_{tr}^{+}} \{Score(d_i)\}$$

$$\tau_N = \max_{d_i \in D_{tr}^{-}} \{Score(d_i)\}$$

A multi-class classification can run in several binary classification. These binary classifications usually have imbalance classification problem, where the number of negative much more than of positive. The performance of imbalance classification problem typically has peak on the left side of data distribution (see Figure 2).

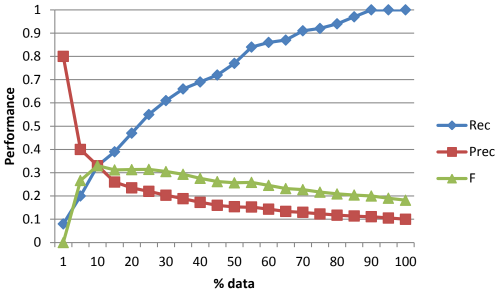


Fig. 2. Typical performance for imbalance classification.

Based on [5] with only score of positive training documents D_{tr}^{+} available, the optimal threshold is τ_P. In a real dataset, in most cases the maximum score of negative testing document are more than the minumum score of positive testing document (see Figure 3). Therefore,

$$\tau = \tau_P - \alpha$$

with simplification version

$$\tau_P \leq \tau \leq \tau_N$$

With both D_{tr}^{+} and D_{tr}^{-} available, we found that

$$\tau = \min(\tau_P, \tau_N)$$

that is

$$\tau = \begin{cases} \tau_P, & \text{if } \tau_P < \tau_N \\ \tau_N, & \text{if } \tau_P > \tau_N \end{cases}$$

5 Evaluation

In this section, we first discuss the data collection used for our experiments. We also describe the baseline models and their implementation.

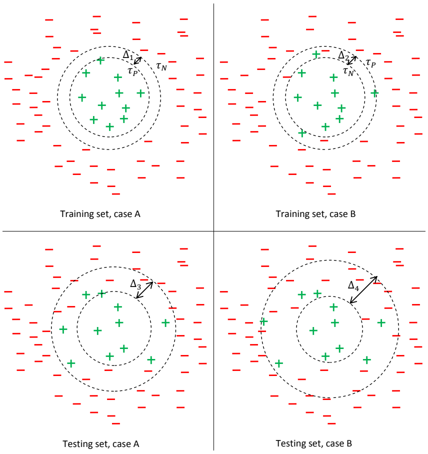


Fig. 3. Training and testing cases. Case A is a non-overlap training score $\tau_P > \tau_N$, case B is an overlap training $\tau_P < \tau_N$. In both case A and case B testing score are overlap, and usually $\Delta_3 < \Delta_4$.

In this study we use the 50 assessor topics of TREC-11 Filtering Track RCV1[1] dataset contains 21,605 documents. According to Buckley and others [1], 50 topics are stable and enough for high quality experiments. RCV1 corpus consists of all and only English language stories produced by Reuter's journalists. Each topic in the dataset is binary class with its own positive and negative set. Positive means that the story is relevant to the assigned topic; otherwise, the word negative will be shown. Documents in TREC-11 are from RCV1, has developed and provided 100 topics for the filtering track aiming at building a robust filtering system. The first 50 topics of TREC-11 RCV1 documents were categorised by humans; the other 50 topics were categorised by intersecting two Reuters topic categories. The assessor topics are more reliable and the quality of the intersection topics is not quite good [10].

Documents in RCV1 are marked in XML. To avoid bias in experiments, all of the meta-data information in the collection have been ignored. The documents are treated as plain text documents by preprocessing the documents. The tasks of removing stop-words according to a given stop-words list and stemming terms by applying the Porter Stemming algorithm are conducted.

Two popular baseline classifiers are used the classic: Rocchio, and SVM. In this paper, our new model is called RFD_τ.

[1] `http://trec.nist.gov/data/t2002_filtering.html`

The Rocchio algorithm [7] has been widely adopted in the area of text categorization. It can be used to build the profile for representing the concept of a topic which consists of a set of relevant (positive) and irrelevant (negative) documents. Two centroids $\boldsymbol{c}_+$ for positive and $\boldsymbol{c}_-$ for negative are generated by using

$$\boldsymbol{c}_+ = \frac{1}{|D^+|} \sum_{\boldsymbol{d} \in D^+} \frac{\boldsymbol{d}}{||\boldsymbol{d}||}$$

$$\boldsymbol{c}_- = \frac{1}{|D^-|} \sum_{\boldsymbol{d} \in D^-} \frac{\boldsymbol{d}}{||\boldsymbol{d}||}$$

For predicting new documents, we use cosine similarity between centroid c_j and document d_i

$$sim(c_j, d_i) = \frac{\boldsymbol{c}_i \cdot \boldsymbol{d}_i}{||\boldsymbol{c}_i|| \times ||\boldsymbol{d}_i||}$$

New documents will be predicted as positive if they are more similar to positive centroid, otherwise it will be predicted as negative.

SVM is a statistical method that can be used to find a hyperplane that best separates two classes. SVM is one of state of the art of text classifier. It achieved the best performance on the Reuters-21578 data collection for document classification [9]. For experiments in this paper, we used SVM^{Light} package [2].

Effectiveness for text classification was measured by two different means F_β and Accuracy (Acc). F_β is the most important metric [9], because F_β is a unification value of Recall (R) and Precision (P):

$$F_\beta = \frac{(\beta^2 + 1)PR}{\beta^2 P + R}$$

The parameter $\beta = 1$ is used in this paper, which means that recall and precision is weighed equally.

$$F_1 = \frac{2PR}{P + R}$$

To get the final result of several topics, two different ways may be adopted, microaveraging (F_1^μ) and macroaveraging (F_1^M) [9]:

$$F_1^\mu = \frac{2P^\mu R^\mu}{(P^\mu + R^\mu)}$$

where

$$P^\mu = \frac{TP}{TP + FP} = \frac{\sum_{i=1}^{|\mathcal{C}|} TP_i}{\sum_{i=1}^{|\mathcal{C}|} (TP_i + FP_i)}$$

$$R^\mu = \frac{TP}{TP + FN} = \frac{\sum_{i=1}^{|\mathcal{C}|} TP_i}{\sum_{i=1}^{|\mathcal{C}|} (TP_i + FN_i)}$$

[2] http://svmlight.joachims.org/

TP (True Positive) is the number of documents the system correctly identifies as positives; TN (True Negative) is the number of documents the system correctly identifies as negatives; FP (False Positive) is the number of documents the system falsely identifies as positives; FN (False Negative) is the number of relevant documents the system fails to identify; and $|\mathcal{C}|$ is the number of topics.

$$F_1^M = \frac{\sum_{i=1}^{|\mathcal{C}|} F_{1,i}}{|\mathcal{C}|}$$

where $F_{1,i}$ is the F_1 for topic i.

For accuracy,

$$Acc = \frac{TP + TN}{TP + FP + TN + FN}$$

$$Acc^{\mu} = \frac{\sum_{i=1}^{|\mathcal{C}|}(TP_i + TN_i)}{\sum_{i=1}^{|\mathcal{C}|}(TP + FP + TN + FN)}$$

$$Acc^M = \frac{\sum_{i=1}^{|\mathcal{C}|} Acc_i}{|\mathcal{C}|}$$

6 Results and Discussion

In this section we present the experimental results for comparing the proposed model with the baseline models. The results listed in Table 2 show that our model outperforms baseline models.

In accuracy, with using original testing for dataset with imbalance number of positive and negative, accuracy produce misleaded measurement. For example in our TREC-11 RCV1 dataset if we predict all of testing documents as negative, we get Recall = 0, Precision=0, F_1 = 0, Accuracy = 81% (macro average), 82% (micro average). If all testing documents are predicted as positive, we get Recall = 1, Precision = 0, F_1 = 0, Accuracy = 19% (macro average) 18% (micro average). Therefore, we use average of five random balance of testing (the number of positive and the number of negative is the same). In balance testing set, if we predict all testing document as negative, we get Recall = 0%, Precision=0%, F_1 = 0, Accuracy $\approx$ 50% (macro average) $\approx$ 50% (micro average). If all testing documents are predicted as positive, Recall = 1, Precision=0, F_1 = 0, Accuracy $\approx$ 50% (macro average) $\approx$ 50% (micro average).

Table 2. Experiment result -with pct change

Model	Macroaverage		Microaverage	
	F_1	Acc	F_1	Acc
RFD_{τ}	**0.43**	**0.68**	**0.53**	**0.70**
SVM	0.19	0.56	0.50	0.61
$Rocchio$	0.37	0.66	0.42	0.68
%change	+16.2%	+3.0%	+6.0%	+2.9%

Table 3. Proportional Thresholding

Model	Macroaverage		Microaverage	
	F_1	Acc	F_1	Acc
RFD_τ	**0.43**	**0.68**	**0.53**	**0.70**
$RFD_{\tau_{prop}}$	**0.43**	0.67	**0.53**	0.69
SVM	0.19	0.56	0.50	0.61
$Rocchio$	0.37	0.66	0.42	0.68

Table 4. p-values for baseline models comparing to RFD_τ model in all accessing topics

	F_1	Acc
SVM	5.349E-12	4.163E-05
$Rocchio$	0.0186	2.555E-08

Table 5. Experiment Result in Scoring Phase [4]

	top-20	*MAP*	F_1	*b/p*	*IAP*
SVM	0.45	0.41	0.42	0.41	0.44
$Rocchio$	0.47	0.43	0.43	0.42	0.45
RFD	**0.56**	**0.49**	**0.47**	**0.47**	**0.51**

For F_1 (Acc), the number of topics where $RFD_\tau > Rocchio$ is 34 (27) topics; and $RFD_\tau > SVM$ is 44 (40) topics.

Compare to proportional thresholding model, an existing popular thresholding model, our thresholding model is comparable (see Table 3). However in proportional thresholding the number of testing dataset has to be known in advance, so it is not suitable for online testing.

The *t-test p* values in Table 4, indicate that the proposed model RFD_τ is statistically significant.

Some classifiers, such as SVM and Rocchio, use scoring then thresholding stages. The final performance of a classifier is based on both stage. In scoring stage, RFD scoring has better result than SVM and Rocchio (see Table 5), where *top-20* is the average precision of the top 20 documents, *MAP* is Mean Average Precision, *b/p* is the break-even point, and *IAP* is Interpolated Average Precision. With an efficient thresholding, scoring methods can maintain its performance.

7 Conclusions

Effective text classification is not a trivial process. In this paper, we proposed a new approach for text classification by using scoring and thresholding. The scoring model is based on pattern mining which capture semantic content of the text. The experiment result for scoring shows that pattern-based scoring is more effective than existing scoring models. While our thresholding model is more powerful than current thresholding model. Furthermore, experiment shows that our whole proposed model as a text classifier is comparable to existing state of the art text classifiers. To improve the performance, for our future work we will break classifier into two parts: recall oriented scoring-thresholding and precision oriented scoring thresholding.

References

1. Buckley, C., Voorhees, E.: Evaluating evaluation measure stability. In: Proceedings of the 23rd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 33–40. ACM (2000)
2. Croft, W., Metzler, D., Strohman, T.: Search engines: Information retrieval in practice. Addison-Wesley (2010)
3. Gopal, S., Yang, Y.: Multilabel classification with meta-level features. In: Proceeding of the 33rd International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 315–322. ACM (2010)
4. Li, Y., Algarni, A., Zhong, N.: Mining positive and negative patterns for relevance feature discovery. In: Proceedings of the 16th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 753–762. ACM (2010)
5. Li, Y., Zhong, N.: Mining ontology for automatically acquiring web user information needs. IEEE Transactions on Knowledge and Data Engineering 18(4), 554–568 (2006)
6. Manning, C., Raghavan, P., Schütze, H.: Introduction to information retrieval, vol. 1. Cambridge University Press, Cambridge (2008)
7. Rocchio, J.: Relevance feedback in information retrieval. In: SMART Retrieval System Experimens in Automatic Document Processing, pp. 313–323 (1971)
8. Schapire, R., Singer, Y., Singhal, A.: Boosting and rocchio applied to text filtering. In: Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 215–223. ACM (1998)
9. Sebastiani, F.: Machine learning in automated text categorization. ACM Computing Surveys (CSUR) 34(1), 1–47 (2002)
10. Soboroff, I., Robertson, S.: Building a filtering test collection for trec 2002. In: Proceedings of the 26th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 243–250. ACM (2003)
11. Wu, S., Li, Y., Xu, Y.: Deploying approaches for pattern refinement in text mining. In: Proceedings of Sixth International Conference on Data Mining, ICDM 2006, pp. 1157–1161 (2006)
12. Yang, Y.: A study of thresholding strategies for text categorization. In: Proceedings of the 24th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 137–145. ACM (2001)
13. Zhang, Y., Callan, J.: Maximum likelihood estimation for filtering thresholds. In: Proceedings of the 24th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 294–302. ACM (2001)

Reference Architectures to Measure Data Completeness across Integrated Databases

Nurul A. Emran[1], Suzanne Embury[2], Paolo Missier[3], and Norashikin Ahmad[1]

[1] Centre of Advanced Computing Technology (C-ACT), University Teknikal Malaysia Melaka
[2] The University of Manchester
[3] The University of Newcastle
{nurulakmar,norashikin}@utem.edu.my,
embury@cs.man.ac.uk, paolo.missier@ncl.ac.uk

Abstract. Completeness is an important aspect of data quality and to determine data acceptability one needs to measure the completeness of the data set of concerned. One type of data completeness measure is population-based completeness (PBC). Nevertheless, the notion of PBC will be of little use until we can determine the efforts required (in terms of architectural design) to implement PBC. In this paper, we present the types of PBC system reference architecture involving integrated databases and motivate the selection of each.

Keywords: reference architecture, population-based completeness (PBC), data completeness, completeness measurement.

1 Introduction

In measuring population-based completeness (PBC) (as proposed in [1]), one essential requirement is to identify and prepare reference population before completeness of data sets of concerned can be measured. Following the definition of PBC, a *simple ratio method* is applied to measure PBC in form of:

$$Completeness(D, RP) = \frac{|(D \cap RP)|}{|RP|} \in [0, 1], \tag{1}$$

where D is the data set under measure, and RP is the reference population.

The situation where there exists a single source that contains a good approximation of the true reference population is limited however (an exception being the Genbank database[1] that contains the genes with good evidence of their existence). Therefore, good approximate populations should be established by integrating individuals from a range of sources. To describe approximate populations established from integrated sources, we adopt the term *universe of discourse (UoD)* or in short *universe*[2]. Conceptually, a universe consists of a collection of approximate populations within an application domain used for PBC

[1] http://www.ncbi.nlm.nih.gov/genbank/

[2] The term *universe* was introduced by Augustus De Morgan in 1846 in formal logics to represent the collection of objects under discussion of a specific discourse [2].

A. Selamat et al. (Eds.): ACIIDS 2013, Part I, LNAI 7802, pp. 216–225, 2013.

measurements, that is built by integrating individuals from several incomplete sources for the populations.

We use the term *contributing sources* (CSs) for sources that contribute to the reference populations in the universe. The CSs could be in multiple forms, such as databases (private and public) e.g., observation databases from gene regulatory networks and pathways [3] or published literature. As it is crucial to understand (and to manage) the relationship between the CSs and the reference populations for successful integration, we propose a structure called a *population map*. Conceptually, a population map consists of a mapping between a reference population and its CSs. If a reference population is stored as a table, and the CSs are databases, we say that a population map is a mapping between a reference population table and queries over tables on CSs. Note that, as the schema of the universe may not be the same as the schema of the CSs, the designers of the population maps must consider the differences.

The notion of PBC and the elements essential to measure PBC will be of little use if they cannot be implemented in practice, in a usable system at 'acceptable cost'. The question of how acceptable the cost is, however, is subjective to the person(s) implementing the PBC system, as it cannot be assessed in absolute terms. Thus, before we can elucidate the costs involved (in terms of software and hardware components), we need to determine the form of PBC system architectures to implement.

The rest of this paper is organised as follows. Section 2 presents the conceptual elements PBC system components; Section 3 covers the forms of PBC reference architecture and Section 4 presents the implementation implications of the reference architectures. Finally, Section 5 concludes the paper.

2 A PBC System

In elucidating the efforts that are involved in developing a PBC system, we need to understand the inputs and processes that are required within a PBC system. The elements of PBC (data set under measure, reference population, universe, CSs) are the inputs of a PBC system, in which the interactions among them are made through some processes that are grouped into three main activities that any PBC implementation must support. These activities are *Setup*, *Measurement* and *Maintenance*. We created a use case diagram to elicit the use cases (that represent the main activities) of a PBC system and the possible interactions among the 'actors'. To elicit the possible interactions among the main activities, we created a main activity diagram showing the plausible flows through them. As the focus of this paper is on Setup, we will present an activity diagram created with the inputs and processes that are specific to it, where the problems and the analysis of their effect on PBC accuracy will be presented in the section to come.

We modelled Setup, Measurement and Maintenance as the main activities that must be supported by a PBC system. These activities are shown as use cases in a use case diagram as shown in Fig. 1. As a result of modelling the main activities of PBC, we identified the 'actors' of the system. These actors are the stakeholders of the system whose interest will enable PBC system to be implemented. The actors and their interactions with the use cases are:

- Data Analyst: the end user of the system who initiates PBC measurement in order to support some data processing tasks that he or she is engaged in. For example, data analysts provide the data set to be measured and the specification of the reference population to use for measurement (e.g., the name of the reference population). This actor might also provide some inputs for the system's maintenance (such as feedback about the performance of the measurement facilities).
- Domain Expert: a person who has knowledge of the populations and the contributing sources of the populations in the application domain, who contributes to Setup and Maintenance.
- System Investor: the person who owns the system and also decides on matters relating to financial cost during Setup and Maintenance.
- PBC Measurement Designer: the person who designs the PBC measures that will be supported by the system and the technical artefacts (e.g., schema) needed to implement them. The PBC Measurement Designer engages in the Setup and Maintenance of the PBC system.
- System Administrator: the person responsible for the maintenance aspect.
- Contributing Source: the source that contributes population individuals in the universe, that interacts during Setup and Maintenance. In some cases, the contributing source might be accessed during Measurement.

We created an activity diagram (Fig. 2) in order to elicit the possible interactions among the main activities, showing the plausible flows through them. Note that [OK] and [Not OK] notations are used to indicate successful and unsuccessful termination of the activities. Among these activities, Setup is the necessary starting point that leads to Measurement and Maintenence. However, if Setup is unsuccessful, due to for example the schema of one or more CSs being inaccessible or the identified CSs no longer existing, we resume Setup until we can deal with the factors that caused the previous attempts to be unsuccessful.

We allow Measurement and Maintenance to occur concurrently as long as the Maintenance does not involve the data sets required by the Measurement. But in the case where the data sets required by the Measurement need to be maintained, the Maintenance needs to wait until the Measurement which is under way is completed. Maintenance must be resumed if it is interrupted for a reason, such as lack of storage space to add new populations.

We have seen so far the actors of the PBC system and the main activities that the actors undertake. We have also described the interactions among the main activities. However, the questions that arise are: 'What is the link among the main activities, the actors and the elements of PBC?', and 'What are the forms of architecture of a PBC system that can be used?' To answer these questions, in the next section, we will present several forms of PBC reference architecture, for which the links among the main activities, the actors and the elements of PBC will be described.

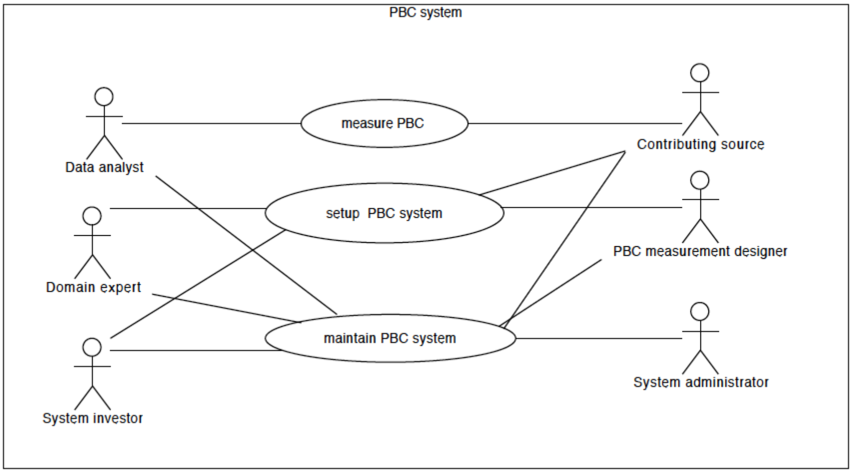


Fig. 1. Main Use Cases of a PBC System

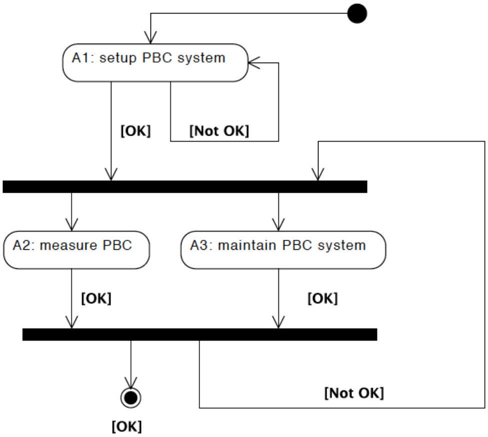


Fig. 2. Main Activities of a PBC system

3 A PBC Reference Architecture

Basically, all PBC system architectures consist of:

- Software components:

  - PBC measurement interface: a software interface for data analysts to issue PBC measurement requests (i.e. by specifying the data set to measure and the specification of the reference population to use) and to receive PBC measurement answers.
  - Measurement processor: a software program that calculates PBC measurement answers based on the PBC measurement requests by instantiating the PBC measurement formula.
  - Wrapper: software that interfaces the universe and the CSs, which is used to transform the queries against CSs that are understandable and executable by the CSs that they wrap.

- Source components:
  - Universe: a database that is either 'virtual' or 'materialised'.
  - Contributing source: source of reference populations in the universe that is in certain forms such as databases (e.g., relational and XML), published literature, or observation data (e.g., data from gene regulatory networks and pathways [TANPI09])).
  - Population map: a structure that stores mappings between populations and their CSs.
  - PBC components configuration: information about the CSs (e.g., CSs name and URI) and the universe (e.g., the reference populations it contains).

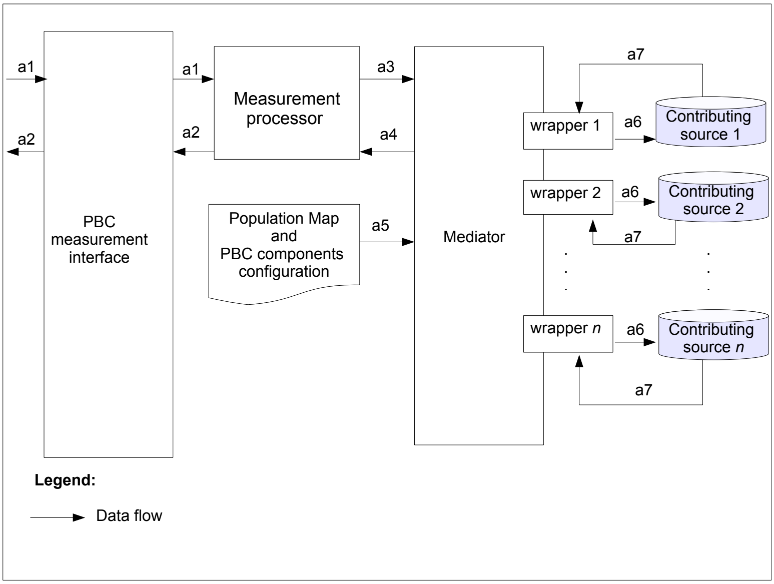


Fig. 3. A PBC Reference Architecture Using a Virtual Universe

The form of the universe selected will determine the type of PCB reference architecture used and the additional software component(s) needed for the architecture.

3.1 The Virtual Universe

We propose the first type of PBC reference architecture as shown in Fig. 3, that implements the universe as a virtual universe. With the virtual universe, it means that the individuals of the reference populations are extracted from the CSs on demand. In addition to the basic components, for this form of architecture, we need a processing layer called a mediator [4,5] that performs data integration tasks on demand. The mediator receives queries for the reference population from the PBC measurement processor (data flow a3), which instantiates the queries based on PBC measurement requests (data flow a1). The mediator executes the queries based on the mappings specified in the population maps and based on CSs information in the PBC components configuration (data flow a5) to extract population individuals from each CS through the wrappers. The queries to extract the populations are shown as data flow a6. These queries are in the form of the schema of the universe (e.g., relational query). Each wrapper will transform the queries into the queries that are understandable by the local CSs (e.g., XML query), which relies on the type of the CSs (e.g., relational or XML databases). The wrappers also transform the query results (population individuals) into the schema that is similar to the universe's schema (called structural transformation). We may rely on the mediator to perform semantic (and format) transformation on the query results to conform to the universe schema. Population individuals from the CSs (data flow a7) are therefore integrated and transformed by the mediator, before they can be passed to the PBC measurement processor for computation (data flow a4). The results of the measurement are sent to the person requesting the measurement (data flow a2).

3.2 The Fully Materialised Universe

Another type of PBC reference architecture that we propose as shown in Fig. 4 implements the universe as a fully materialised database (whose storage approach is similar to a classic datawarehouse design [6]). Instead of querying reference populations on an on-demand basis, the individuals of the reference populations that are extracted from the CSs (data flow a7) are stored within the universe. A similar architecture is adopted by a yeast datawarehouse for datamining purposes [7]. For this type of architecture, we need a software that could support up-front extraction of population individuals from multiple CSs, prior to the actual PBC measurements. The extraction process relies on the mappings between the populations and their CSs (retrieved from the population maps), and the information about PBC components configuration (data flow a5). The extracted data sets must also conform to the schema of the universe, before they can be loaded into the universe. Therefore, in addition to the basic components, we propose to include an ETL[3] pipeline within the architecture that will support the tasks just mentioned. The wrappers are plugged into the ETL pipeline and the functionality of the wrappers in this architecture is the same as the wrappers in virtual universe reference architectures. The Transform component is responsible for performing semantic (and formatting) transformation and for integrating the population individuals, which is the same functionality that is performed by the mediator. The Load component loads the integrated reference populations into the universe, through the universe management interface. In this architecture, we also propose to include a software interface, called the universe management interface so that system administrators can interact with the ETL pipeline e.g., in receiving the transformed population individual data sets (data flow a8). Information about modifications on the CSs (data flow a9) is needed in order to perform updates on the universe.

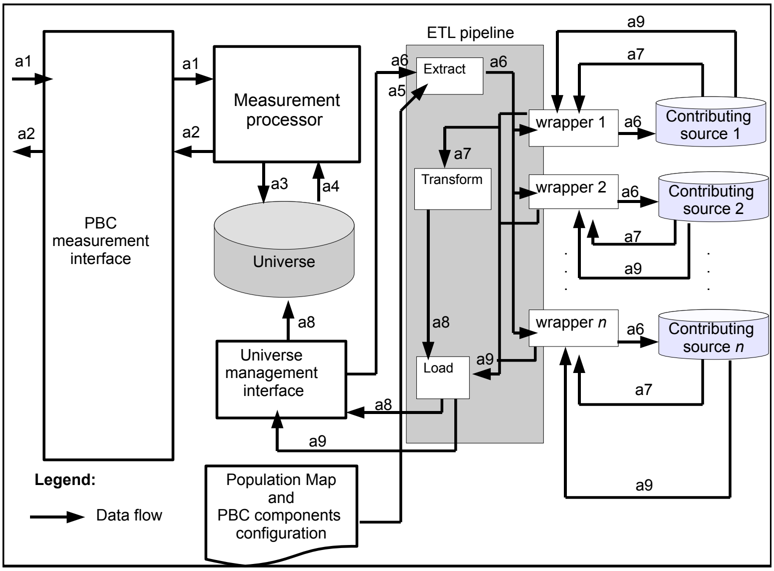


Fig. 4. A PBC Reference Architecture Using a Fully Materialised Universe

3.3 The Partially Materialised Universe

A reference architecture that implements a partially materialised universe, known as a hybrid approach [9,10], as shown in Fig. 5, is another type of architecture that is plausible for a PBC system. In this architecture, only some reference populations are stored in the universe while the others are kept in their own CSs and queried on demand through a mediator. A PBC reference architecture with

[3] Extract, transform and load. A software tool that is usually used in data warehousing and decision support system [8].

a partially materialised universe must support queries for reference populations (data flow a3) against both the universe and the CSs. We propose to include a software component in this architecture, called the 'broker', that will decide whether queries for the reference population should be sent to the universe or to the mediator (or both). The broker must know which reference populations are materialised and which must be fetched on demand. The measurement processor will interact with the broker in issuing the queries instantiated based on the PBC measurement requests. Data flows involving the virtual universe (through the mediator) are illustrated in blue for a clearer distinction between the functionalities of fully materialised and virtual universe architecture within this hybrid architecture.

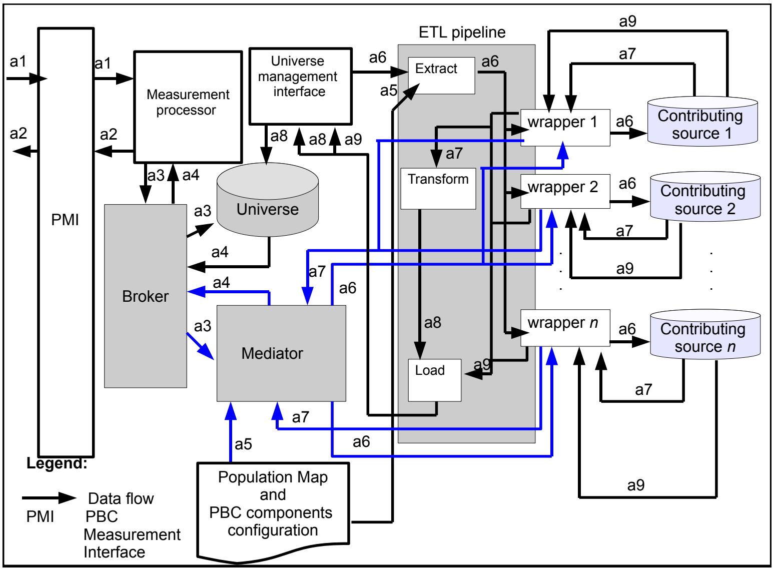


Fig. 5. A PBC Reference Architecture Using a Partially Materialised Universe

4 Implementation Implications

PBC measurement providers might choose one particular form of PBC reference architecture to implement PBC for a number of reasons.

The architecture with the virtual universe might be chosen because the reference populations are queried on-demand; therefore actual load of data is unnecessary except during the measurement. As a consequence, small data storage allocation is needed to store the system's components configuration information (such as for the CSs configuration) and population maps. However, the virtual universe is unlikely to be very efficient if PBC measurement needs to be performed frequently, and the number of CSs that must be queried to perform the measurement is large. This is because there is a risk of communication failures, especially if the CSs are prone to experiencing technical problems that cause restricted access. In addition, PBC measurement is likely to be slow, as all the complicated tasks of querying the population individuals from multiple CSs, transforming the query results to conform with the universe schema, and combining the query results must take place before the measurement processor can initiate the PBC computation.

PBC reference architecture with a fully materialised universe serves as a better alternative if quick PBC measurement is needed. Even though an up-front effort to construct and populate the universe is needed, it could be worthwhile if a large number of data analysts are accessing the reference populations, each wishing to make use of a slightly different combination of the data set under measure and the reference population. However, we need to deal with data freshness, an issue that is common for any approach that materialises copies of data sets from other data sources. Therefore, it is necessary to put effort into detecting or monitoring updates to the contributing data sources, to ensure the materialised data sets are up to date [11].

With additional knowledge regarding which reference populations are frequently accessed (or are not frequently accessed), one might consider PBC reference architecture that implements a partially materialised universe. In terms of the requirements, a partially materialised universe needs support for the functionality required by a fully materialised universe and a virtual universe.

Regardless of the reasons for choosing a particular type of PBC reference architecture, the components in all architectures have their own functionality to support PBC measurement providers to answer PBC measurement requests. The data flows in the architectures are present as the result of the interactions among the components of the system (software and source) that each serve their own functionality (e.g., measurement processor to accept PBC request, instantiate the query and pass it to the universe/mediator). In comparison, with the fully materialised universe, a PBC reference architecture consists of all data flows that are present in the architecture that implements the virtual universe (data flow a1 to a6) with two extra additional data flows (data flow a7 and a8); all data flows that are present in the partially materialised universe architecture can be found in the fully materialised universe architecture. Using the data flows to indicate the results of the functionalities performed by the components, we say that the differences (in terms of functionality) among the three forms of architecture are small.

5 Conclusion

In this paper, we proposed three types of PBC reference architecture, each with extra components that must be added along with the basic components of PBC. The choice of the type of PBC reference architecture to use will guide PBC measurement providers to determine the software (and hardware) components that are needed and therefore to make necessary estimates on the financial costs incurred. However, the question that remains unanswered is regarding the efforts needed to support PBC measurement providers in providing accurate PBC measurement answers. In our future work, implementation of these framework will be done to validate the hypothesized implementation implications.

Acknowledgments. The authors would like to thank the financial assistance provided by the Universiti Teknikal Malaysia, Melaka (UTeM) and the Ministry of Higher Education, Malaysia during the course of this research.

References

1. Emran, N.A., Embury, S., Missier, P., Muda, A.K.: Measuring Data Completeness for Microbial Genomics Database. In: Selamat, A., et al. (eds.) ACIIDS 2013, Part I. LNCS (LNAI), vol. 7802, pp. 186–195. Springer, Heidelberg (2013)
2. Maddux, R.: The origin of relation algebras in the development and axiomatization of the calculus of relations. Studia Logica 50, 421–455 (1991)
3. Tiffin, N., Andrade-Navarro, M.A., Perez-Iratxeta, C.: Linking genes to diseases: it's all in the data. Genome Medicine 1, 1–7 (2009)
4. Lenzerini, M.: Data integration: a theoretical perspective. In: Proceedings of the Twenty-first ACM SIGMOD-SIGACT-SIGART Symposium on Principles of Database Systems (PODS), pp. 233–246. ACM (2002)
5. Wiederhold, G.: Mediators in the architecture of future information systems. Computer 25, 38–49 (1992)
6. Inmon, W.: In: Elliot, R. (ed.) Building the Datawarehouse, 3rd edn., pp. 1–427. Wiley Computer Publishing (2002)
7. Balakrishnan, R., Park, J., Karra, K., Hitz, B., Binkley, G., Hong, E., Sullivan, J., Micklem, G., Cherry, J.: YeastMinean integrated data warehouse for Saccharomyces cerevisiae data as a multipurpose tool-kit. Database: The Journal of Biological Database and Curation 2012, 1–8 (2012)
8. Chaudhuri, S., Dayal, U., Ganti, V.: Database technology for decision support systems. Computer 34, 48–55 (2001)
9. Hull, R., Zhou, G.: A framework for supporting data integration using the materialized and virtual approaches. SIGMOD Records 25, 481–492 (1996)
10. Hull, R.: Managing semantic heterogeneity in databases: a theoretical prospective. In: Proceedings of the Sixteenth ACM SIGACT-SIGMOD-SIGART Symposium on Principles of Database Systems (PODS), pp. 51–61. ACM (1997)
11. Widom, J.: Research problems in data warehousing. In: Proceedings of the Fourth International Conference on Information and Knowledge Management, pp. 25–30. ACM (1995)

Vietnamese Author Name Disambiguation for Integrating Publications from Heterogeneous Sources

Tin Huynh, Kiem Hoang, Tien Do, and Duc Huynh

University of Information Technology, Vietnam,
Km 20, Hanoi Highway, Linh Trung Ward, Thu Duc District, HCMC
{tinhn,kiemhv,tiendv}@uit.edu.vn,
duc2802@gmail.com

Abstract. Automatic integration of bibliographical data from various sources is a really critical task in the field of digital libraries. One of the most important challenges for this process is the author name disambiguation. In this paper, we applied supervised learning approach and proposed a set of features that can be used to assist training classifiers in disambiguating Vietnamese author names. In order to evaluate efficiency of the proposed features set, we did experiments on five supervised learning methods: Random Forest, Support Vector Machine (SVM), k-Nearest Neighbors (kNN), C4.5 (Decision Tree), Bayes. The experiment dataset collected from three online digital libraries such as Microsoft Academic Search[1], ACM Digital Library[2], IEEE Digital Library[3]. Our experiments shown that kNN, Random Forest, C4.5 classifier outperform than the others. The average accuracy archived with kNN approximates 94.55%, random forest is 94.23%, C4.5 is 93.98%, SVM is 91.91% and Bayes is lowest with 81.56%. Summary, we archived the highest accuracy 98.39% for author name disambiguation problem with the proposed feature set in our experiments on the Vietnamese authors dataset.

Keywords: Digital Library, Data Integration, Bibliographical Data, Author Disambiguation, Machine Learning.

1 Introduction

In our previous work, we proposed and developed a system that is used 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 [9]. When we integrate information from various heterogeneous information sources, we must identify data records that refer to equivalent entities. One of the most important challenges for this problem is the author name disambiguation.

[1] http://academic.research.microsoft.com

[2] http://dl.acm.org/

[3] http://www.ieeexplore.ieee.org/

A. Selamat et al. (Eds.): ACIIDS 2013, Part I, LNAI 7802, pp. 226–235, 2013.

Table 1. The illustrative example, papers contained ambiguous author names

Paper no.1	*Multiagent Place-Based Virtual Communities for Pervasive Computing.* Conference: PERCOM '08 Proceedings of the 2008 Sixth Annual IEEE International Conference on Pervasive Computing and Communications Authors: ***Tuan Nguyen,*** Seng Loke; Torabi, T.; Hongen Lu. Dept. of Comput. Sci. & Comput. Eng., La Trobe Univ., Bundoora, VIC.	Mr. Tuan worked at Dept. of Comput. Sci. & Comput. Eng., La Trobe Univ., VIC., AU.
Paper no.2	*Stationary points of a kurtosis maximization algorithm for blind signal separation and antenna beamforming.* Journal: Journal IEEE Transactions on Signal Processing Authors: Zhi Ding, ***Tuan Nguyen*** Dept. of Electr. & Comput. Eng., Iowa Univ., Iowa City, IA.	Mr. Tuan worked at Dept. of Electr. & Comput. Eng., Iowa Univ., Iowa City, IA.
Paper no.3	*Semantic-PlaceBrowser: Understanding Place for Place-Scale Context-Aware Computing* Conference: The Eighth International Conference on Pervasive Computing (Pervasive 2010), Helinski, Finland, 2010. Authors: ***Anh-Tuan Nguyen***, Seng Wai Loke, Torab Torabi, Hongen Lu. Department of Computer Science and Computer Engineering, La Trobe University, Victoria, 3086, Australia	Mr. Tuan worked at Dept. of Comput. Sci. & Comput. Eng., La Trobe Univ., VIC., AU.

Name ambiguity is a problem that occurs when a set of publications contains ambiguous author names, i.e the same author may appear under distinct names (synonyms), or distinct authors may have similar names (polysems) [3]. The table 1 shows an example about Vietnamese author name ambiguity in 3 different publications. Author names in the paper 1 and the paper 3 are examples of synonyms. Both refer to 'Tuan Nguyen' from Department of Computer Science & Computer Engineering, La Trobe University, Melbourne, Victoria Australia. While author names in the paper 1 and the paper 2 are examples of polysems where 'Tuan Nguyen' in the paper 1 refers to 'Tuan Nguyen' from Department of Computer Science & Computer Engineering, La Trobe University, Melbourne, Victoria Australia and 'Tuan Nguyen' in the paper 2 refers to 'Tuan Nguyen' from the Department of Electric & Computer Engineering, Iowa University, Iowa City, IA, United States.

Name ambiguity in this context is a critical problem that have attracted a lot of attention of the digital library research community. Especially, authors with Asian names are very ambiguous cases that still be one of open challenges for this problem [3]. Therefore, we mainly focus on Vietnamese author name disambiguation for integrating the bibliographical data.

In section 2, we briefly present related research on developing digital libraries, building bibliographical database, disambiguating automatically author name. Section 3 presents our set of features that can be used to assist author name disambiguating in the context of bibliographical data integration of computer science publication from various online digital libraries. The experiments, evaluation and discussion will be introduced in section 4. We conclude the paper and suggest future works in section 5.

2 Related Work

Ferreira et al [3] did a brief survey of automatic methods for author name disambiguation. For this problem, the proposed methods usually attempt to group citation records of a same author by finding some similarity among them (author grouping methods) or try to directly assign them to their respective authors (author assignment methods). Both approaches may either exploit supervised or unsupervised techniques. They also reviewed classification, clustering algorithms and as well as similarity functions which can be applied for this problem.

Han et al. applied a model based k-means clustering algorithm and a K-way spectral clustering algorithm to solve this problem [6][7]. Beside that, Han et al. also explored two supervised learning algorithms, Naive Bayes and Support Vector Machine. Both these supervised algorithms uses three types of citation attributes coauthor names, paper title keywords and journal title keywords, and achieved more than 90.0% accuracies in disambiguation [5]. In another research, Huang et al. presented an efficient integrative framework for solving the name disambiguation problem: a blocking method retrieves candidate classes of authors with similar names and a clustering method, DBSCAN, clusters papers by author [8]. For evaluation, they manually annotated 3,355 papers yielding 490 authors and achieved 90.6% for average accuracy.

Treeratpituk and Giles [11] also proposed random forest algorithm and a feature set for disambiguating author names in Medline dataset, a bibliographic database of life sciences and biomedical information. They applied and compared the random forests with some other algorithms such as NaiveBayes, Support Vector Machine, Decision Tree in this problem. Their experiments showed that the random forest outperforms than the others with the proposed feature set. The highest accuracy archived with random forest is 95.99% for the Medline dataset and their proposed feature set.

Qian et al proposed a labeling oriented author disambiguation approach, called LOAD, to combine machine learning and human judgment together in author disambiguation [10]. Their system have a feature that allows users to edit and correct author's publication list. All such user edited data are collected as UE (User Edited) dataset. They did experiments on their the UE dataset and DBLP dataset. The average accuracy archived 91.85% with their proposed method.

In general, author assignment methods mainly focused on learning a model for each specific author. They are usually very effective when faced with a large number of examples of citations for each author. However, DLs are very dynamic systems, thus manual labeling of large volumes of examples is unfeasible. In addition, authors often change their interesting area over time, new examples need be insert into training data continuously and the methods need to be retrained periodically in order to maintain their effectiveness [3].

While author grouping methods mainly focused on defining a similarity function used to group corresponding publications using clustering technique. The similarity function may be predefined or learned by using supervised learning methods. Learning a specific similarity function usually produces better results [3]. Therefore, in this research we proposed a set of features for learning a similarity function by using supervised learning methods.

3 Our Approach

How we can disambiguate these ambiguous names in two different publications. In order to do that, we can base on similarity of metadata of these publications. For example, similarity of names, affiliations, list of coauthors, keywords in publications which contain these ambiguous authors. Based on these metadata, we applied supervised learning methods and proposed a set of features that can be used to support training classifiers or learning a similarity function.

In this section, we present the popular string matching methods and consider to apply them for computing the similarity of metadata. We also present the proposed feature set for learning a similarity function.

3.1 Popular String Matching Measures

In [1][2], authors reviewed widely used measures in measuring similarity of two strings to identify the duplication. These measures basically are divided into three categories: (1) *Edit distance*; (2) *Token-based* and hybrid methods.

Edit Distance: distance between strings X and Y is the cost of the best sequence of *edit operations* that converts X to Y. There are some edit distance measures such as Levenshtein, Monger-Elkan, Jaro, Jaro-Winkler [2]. Levenshtein is one of the most popular measure for edit distance. For Levenshtein, the cost of converting X to Y is computed by three types of edit operations: (1) inserting a character into the string; (2) deleting a character into the string and (3) replacing one character with a different character.

$$Sim_{levenshtein}(X,Y) = 1 - \frac{d(X,Y)}{[max(length(X), length(Y))]}$$

Where:

- $d(X,Y)$ the minimum number of edit operation of single characters needed to transform the string X into Y.
- $length(x)$ the length of string X.

Token-Based Measures: in many situations, word order is not really important. In such cases, we can convert the strings X and Y to token multisets and consider similarity metrics on these multisets. Jaccard, TF/IDF [2], popular token based measures, widely used.

Hybrid Measures: Mogne-Elkan measure propose the following recursive matching scheme for comparing two long strings X and Y [2]. First, X and Y are broken into substrings $X = x_1...x_K$ and $Y = y_1...y_L$. Then, similarity is defined as:

$$Sim_{monge-elkan}(X,Y) = \frac{1}{K}\sum_{i=1}^{K}\max_{j=1}^{L} Sim^{'}(x_i, y_j)$$

Where:

- $Sim^{'}$: is some secondary Edit Distance as Levenshtein, Jaro-Winkler.

In our problem, similarity of metadata strings (author name, co-authors, affiliation, keywords in publications) mainly based on tokens. Especially, most of ambiguous Vietnamese author names relate to the order of their family name and given name. The first order in full name sometimes is family name, while other cases are given name. Therefore, we have applied Jaccard method to calculate the similarity for features in our proposed feature set.

3.2 The Proposed Feature Set

To learn a similarity function, the disambiguation methods receive a set of pais of publications as the training dataset. The similarity of two ambiguous authors in a pair of publications is presented by a multidimensional vector. Therefore, we are going to pair publications that relate to ambiguous authors in our dataset. After that we encode similarity of this pair by a vector. In our dataset, for each pair of publications we labeled true (1) if two ambiguous authors are the same person or false (0) if two ambiguous authors are two different persons. We applied supervised learning algorithms to train, test and evaluate the performance of the proposed feature set. The detail of proposed feature set is described as following:

Author Name Similarity: We can assume that if the similarity of two string used to presented names of two ambiguous authors is high then two these ambiguous authors may relate to the same person. In order to calculate the similarity of names of these ambiguous authors, we used Jaccard coefficient.

$$Author\_Name\_Sim(A,B) = \frac{|Author\_Name\_A \bigcap Author\_Name\_B|}{|Author\_Name\_A \bigcup Author\_Name\_B|}$$

Where:

- $Author\_Name\_A$: name of author A presented in one specified publication.
- $|Author\_Name\_A \bigcap Author\_Name_B|$: number of same tokens in names of A and B.
- $|Author\_Name\_A \bigcup Author\_Name\_B|$: number of tokens in Author_Name_A or Author_Name_B.

For example:

Author_Name_A="Tuan Nguyen" and Author_Name_B="Nguyen Anh Tuan"
$|Author\_Name\_A \bigcap Author\_Name\_B| = 2$
$|Author\_Name\_A \bigcup Author\_Name\_B| = 3$
and Author_Name_Sim(A, B) = 0,6.

Affiliation Similarity: Affiliations of two ambiguous authors in two different publications is one of feature that can used to recognize whether two ambiguous authors actually mention to one person. In order to calculate the similarity of names of these affiliations, we also used Jaccard coefficient.

$$Aff\_Sim(A,B) = \frac{|Aff\_A \bigcap Aff\_B|}{|Aff\_A \bigcup Aff\_B|}$$

Where:

- Aff_A: Affiliation name of author A in one specified publication
- $|Aff\_A \bigcap Aff\_B|$: number of same tokens between affiliation of author A and affiliation of author B.
- $|Aff\_A \bigcup Aff\_B|$: number of tokens in string presented affiliations name of author A or author B.

CoAuthors Similarity: If two ambiguous author names share at least one same coauthor in two different publications then these ambiguous names may be one author. We proposed calculating method for the Coauthors_Name_Sim feature as following:

$$CoAuthors\_Names\_Sim(A, B) = MAX(Author\_Name\_Sim(A_i, B_j))$$

Where:

- $A_i \in CoAuthors(A)$
  and CoAuthors(A): is a set of co-authors of author A in publication P_1.
- $B_j \in CoAuthors(B)$
  and CoAuthors(B): is a set of co-authors of author B in publication P_2.

CoAuthor_Affs Similarity: If two ambiguous author names have coauthors who have worked in the same university or institute then these ambiguous names may be one author. We proposed calculating method for feature based on affiliation of coauthors as following:

$$CoAuthor\_Affs\_Sim(A, B) = MAX(Aff\_Sim(Aff_i\_P1, Aff_j\_P_2))$$

Where:

- $Aff_i\_P1 \in CoAuthors\_Affs(A)$
  i=1..n (n: number of coauthor of author A in P_1)
  and CoAuthors_Affs(A): is a set of affiliations of co-authors of author A in publication P_1.
- $Aff_i\_P2 \in CoAuthors\_Affs(B)$
  j=1..m (m: number of coauthor of author B in P_2)
  and CoAuthors_Affs(B): is a set of affiliations of co-authors of author B in publication P_1.

Paper_Keywords Similarity: If publications related to ambiguous authors contain similar keywords then two ambiguous authors may be the same person. We proposed the calculating method for the keyword-based feature as following:

$$Paper\_Keywords\_Sim(A, B) = \frac{|Paper\_Keywords\_A \bigcap Paper\_Keywords\_B|}{|Paper\_Keywords\_A \bigcup Paper\_Keywords\_B|}$$

Where:

- $Paper\_Keywords\_A$: is the set of keywords in publication A (publication of author A).
- $|Paper\_Keywords\_A \bigcap Paper\_Keywords_B|$: number of same tokens which keywords in publication A and publication B share.
- $|Paper\_Keywords\_A \bigcup Paper\_Keywords\_B|$: number of tokens of keywords in publication A or B.

4 Experiments and Evaluation

In order to analyze how classifiers and the feature set performed for the author name disambiguation problem, we applied the proposed feature set to many different classifiers such as Random Forest, kNN, SVM, C4.5, Bayes Nets. We did experiments to consider which classifier is to bring out higher accuracy and how classifiers and the proposed feature set effect to the author name disambiguation. This section presents our experimental results and discussions on archived results.

4.1 Dataset

Experimental dataset collected from three online digital libraries that are ACM DL, IEEE Xplore, MAS. In order to check the author name ambiguity when integrating publications from many various sources, we prepared 10 author names as the input data that used to submit to these digital libraries. All of these author names are Vietnamese author names (very ambiguous cases). For these authors, there are many different instance names. For example, author 'Kiem Hoang' can have many different names in the database such as 'Hoang Kiem', 'Kiem Hoang', 'Hoang Van Kiem', 'Kiem Van Hoang'.

Table 2. The dataset collected and labeled for training and testing

Submitted names	Number of publications collected	Number of pairs labeled	Number of pairs labeled with value 0	Number of pairs labeled with value 1
Cao Hoang Tru	30	435	26	409
Dinh Dien	30	435	51	384
Duong Anh Duc	30	435	29	406
Ha Quang Thuy	30	435	84	351
Ho Tu Bao	30	435	0	435
Kiem Hoang	30	435	25	410
Le Dinh Duy	30	435	57	378
Le Hoai Bac	30	435	0	435
Nguyen Ngoc Thanh	30	435	141	294
Phan Thi Tuoi	30	435	0	435
Total	**300**	**4350**	**413≅9.49%**	**3937≅90.51%**

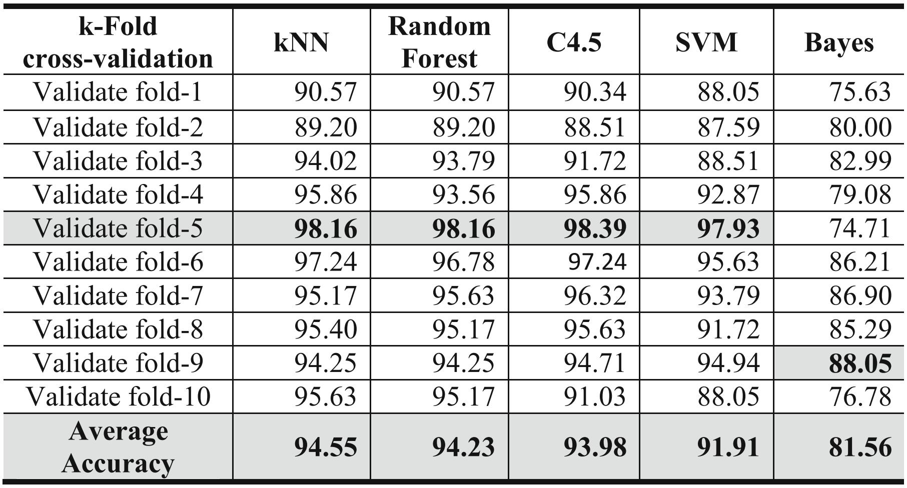
For each author in this set, we collected 30 publications returned from 3 these online libraries. We do not care about the duplicate of publication from these libraries. These publications contain ambiguous author names for the integration (two different authors with the same name or one author with different names). We built the training and testing dataset by paring of publications in 30 publications for each author. Based on our understanding about these authors, we labeled for each pair with value 1 if ambiguous names in this pair actually be one person and value 0 if these ambiguous names actually be two different persons. So, there are totally 4350 samples in our dataset. The table 2 show the detail of the dataset. Each sample in the dataset relate to a pair of publications which contain ambiguous authors. Each sample will be presented by a vector its each dimension is one specified feature.

4.2 Experiments for author disambiguation

We applied the k-fold cross validation, a method checks how well a model generalizes to new data, to evaluate the proposed feature set and various classifiers in our experiments. We used different supervised learning algorithms implemented by the WEKA project [4]. At this time, we tested with 5 various classifiers in WEKA such as kNN, Support Vector Machine, Random Forest, C4.5, Bayes Network. There are totally 4350 samples in the dataset.

The dataset is divided into 10 subsets with the same size. The cross-validation process is then repeated 10 times (folds). Each time, one of 10 subsets sequentially is used as test set and the remaining subsets are put together to form a training set. Therefore, each subset is used exactly once as the validation data. The 10 results from 10 folds then can be averaged to produce a estimation or an evaluation for classifiers. Experimental results, showing the accuracy of author name disambiguation with the proposed feature set and classifiers, are reported in table 3.

Table 3. The experimental results and evaluations with k-fold validation



k-Fold cross-validation	kNN	Random Forest	C4.5	SVM	Bayes
Validate fold-1	90.57	90.57	90.34	88.05	75.63
Validate fold-2	89.20	89.20	88.51	87.59	80.00
Validate fold-3	94.02	93.79	91.72	88.51	82.99
Validate fold-4	95.86	93.56	95.86	92.87	79.08
Validate fold-5	**98.16**	**98.16**	**98.39**	**97.93**	74.71
Validate fold-6	97.24	96.78	97.24	95.63	86.21
Validate fold-7	95.17	95.63	96.32	93.79	86.90
Validate fold-8	95.40	95.17	95.63	91.72	85.29
Validate fold-9	94.25	94.25	94.71	94.94	**88.05**
Validate fold-10	95.63	95.17	91.03	88.05	76.78
Average Accuracy	**94.55**	**94.23**	**93.98**	**91.91**	**81.56**

4.3 Discussion

Figure 1 compares the accuracy using of 5 classifiers based on the proposed feature set for the author name disambiguation problem. The figure presented experimental results for all 10-folds cross-validation. The average accuracy of these 10-folds is shown in table 3. Our experiments shown that kNN, Random Forest, C4.5 classifier outperform than the others in most of folds (figure 1). The average accuracy archived with kNN approximates 94.55%, random forest is 94.23% and C4.5 is 93.98%. While the accuracy of SVM algorithm is 91.91%. Bayes classifier give lowest accuracy with this problem. The average accuracy archived with Bayes classifier approximates 81.56%. Summary, we archived the highest accuracy 98.39% for the author name disambiguation problem with our proposed feature set (table 3).

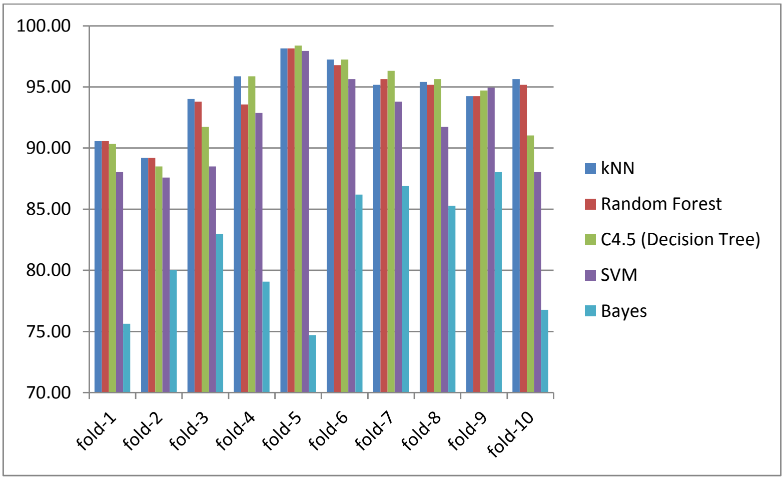


Fig. 1. The various accuracies produced by 10-folds cross-validation for author name disambiguation using different classifiers

5 Conclusion and Future Work

The gold of this research is to evaluate the performance of the proposed feature set which can be used to support training classifiers for disambiguating author names. Author name disambiguation is really a critical task for integrating computer science publications from various sources. We proposed and presented the feature set based on metadata that can be used to disambiguate author names in different publications. The feature set tested with 5 various classification algorithms. We built the dataset for training and testing based on publications contain Vietnamese authors. Because most of Vietnamese authors who often apply many different way to write names in their publications. The results show that our proposed feature set archived the average accuracy is 94.55% with kNN and the highest accuracy is 98.39% with the C4.5 algorithm (table 3).

In the future, we are going to continue improving the accuracy for this problem and doing more experiments for asian author name disambiguation and other very ambiguous cases on DBLP, a public dataset. We will combine this module of author names disambiguation with our existing system that used to integrate computer science publications from various online sources.

Acknowledgments. This research is partially supported by VNU-HCM R&D project: B2011-26-01.

References

1. Bilenko, M., Mooney, R., Cohen, W., Ravikumar, P., Fienberg, S.: Adaptive name matching in information integration. IEEE Intelligent Systems 18(5), 16–23 (2003)
2. Cohen, W.W., Ravikumar, P.D., Fienberg, S.E.: A comparison of string distance metrics for name-matching tasks. In: IIWeb, pp. 73–78 (2003)
3. Ferreira, A.A., Gonçalves, M.A., Laender, A.H.: A brief survey of automatic methods for author name disambiguation. SIGMOD Rec. 41(2), 15–26 (2012)
4. Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., Witten, I.H.: The weka data mining software: an update. SIGKDD Explor. Newsl. 11(1), 10–18 (2009)
5. Han, H., Giles, L., Zha, H., Li, C., Tsioutsiouliklis, K.: Two supervised learning approaches for name disambiguation in author citations. In: Proceedings of the 4th ACM/IEEE-CS Joint Conference on Digital Libraries, JCDL 2004, pp. 296–305. ACM, New York (2004)
6. Han, H., Zha, H., Giles, C.L.: A model-based k-means algorithm for name disambiguation. In: Proceedings of Semantic Web Technologies for Searching and Retrieving Scientific Data, Florida, USA (October 20, 2003)
7. Han, H., Zha, H., Giles, C.L.: Name disambiguation in author citations using a k-way spectral clustering method. In: Proceedings of the 5th ACM/IEEE-CS Joint Conference on Digital Libraries, JCDL 2005, pp. 334–343. ACM, New York (2005)
8. Huang, J., Ertekin, S., Giles, C.L.: Efficient Name Disambiguation for Large-Scale Databases. In: Fürnkranz, J., Scheffer, T., Spiliopoulou, M. (eds.) PKDD 2006. LNCS (LNAI), vol. 4213, pp. 536–544. Springer, Heidelberg (2006)
9. Huynh, T., Luong, H., Hoang, K.: Integrating Bibliographical Data of Computer Science Publications from Online Digital Libraries. In: Pan, J.-S., Chen, S.-M., Nguyen, N.T. (eds.) ACIIDS 2012, Part III. LNCS, vol. 7198, pp. 226–235. Springer, Heidelberg (2012)
10. Qian, Y., Hu, Y., Cui, J., Zheng, Q., Nie, Z.: Combining machine learning and human judgment in author disambiguation. In: Proceedings of the 20th ACM International Conference on Information and Knowledge Management, CIKM 2011, pp. 1241–1246. ACM, New York (2011)
11. Treeratpituk, P., Giles, C.L.: Disambiguating authors in academic publications using random forests. In: Proceedings of the 9th ACM/IEEE-CS Joint Conference on Digital Libraries, JCDL 2009, pp. 39–48. ACM, New York (2009)

Retrieval with Semantic Sieve

Julian Szymański, Henryk Krawczyk, and Marcin Deptuła

Faculty of Electronics, Telecommunications and Informatics,
Department of Computer Systems Architecture,
Gdańsk University of Technology, Poland
{julian.szymanski,henryk.krawczyk}@eti.pg.gda.pl,
marcin.deptula@live.com

Abstract. The article presents an algorithm we called *Semantic Sieve* applied for refining search results in text documents repository. The algorithm calculates so-called *conceptual directions* that enables interaction with the user and allows to narrow the set of results to the most relevant ones. We present the system where the algorithm has been implemented. The system also offers in the presentation layer clustering of the results into thematic groups. Preliminary evaluation indicates the proposed approach can be useful for precessing search results and serve as effective tool for improving retrieval with keywords.

Keywords: information retrieval, documents clustering, text processing.

1 Introduction

The information in World Wide Web is mostly stored in the form of a text written in a natural language. Excluding the multimedia data, WWW is a dynamic repository of text documents where searching for relevant information is done with providing to the user web pages (text documents) that fulfill his or her requirements. The specification of the user requirements for retrieving the textual information is mainly carried out with the usage of keyword-based methods. The typical interaction with the search engine is a scenario in which the user provides poorly defined search criteria (given as search phrases) and expects to receive a relevant information.

Keyword-based searches are known to be helpful for retrieval of a relevant content from large text collections. They are especially useful when strengthened with the results ranking, but the keyword based retrieval has some serious limitations: first of all the ambiguity of natural language causes that the same content can be expressed in many ways. The ambiguity appears not only at the level of sentences and expressions but even at words level which can be seen in the existence of synonyms and homonyms. The user, while retrieves the content with keyword-based methods needs to know exact phrases that index the text documents which are interesting for his or her requirements.

The improvements of keyword-based retrieval can be performed in many ways, e.g. by adding to search phrases their synonyms that extend the set of indexed content. This task is realized by using semantic dictionaries or based on the results of statistical analysis of large text collections. In that case the *recall* of the search increases but the *precision* degradates [1] [2].

A. Selamat et al. (Eds.): ACIIDS 2013, Part I, LNAI 7802, pp. 236–245, 2013.

For the search improvement taxonomies of concepts are also useful they allow to capture generalizations of words. The other way for improving keyword-based retrieval are the similarity based searches that find a content that is similar, but not necessary contains search phrases. The ambiguity of natural language causes that the keywords do not index all relevant information. On the other hand it is also typical that the user doesn't know proper keywords that index relevant content to his or her requirements. If he or she doesen't know the exact keywords, retrieving by using descriptions of required results is needed. This approach encounters some difficulties, especially resulting in loosing the search precision.

The other issue is the fact that while searching we often even don't know precisely what we want to find until we see the retrieved results. So during search process the suggestions of a content that can be interested to the user is also valuable. These suggestions may be realized in many ways, but one of the most useful is propose the content that is conceptually related to the defined search criteria, not only to the pages exactly matching the searched phrase.

Thus the amount of the data stored in documents in a natural language requires advanced methods for information retrieval [3] and effective mechanisms to extract knowledge from the text. The main barrier to achieve this is that the machine does not understand the human language and thus all improvements in information retrieval can be made only with improving the user queries interpretation.

In this paper we present the algorithm for retrieval of documents that is based on interaction with the user. The algorithm has been applied for Wikipedia articles where it extends a keyword-based retrieval and allows the user to precise his search requirements which narrow the result set to the more relevant content.

2 Semantic Sieve Algorithm

Think about something - I'll try to guess it. This is an idea of popular word game where one attendant is thinking about something and the other is asking the questions and tries to guess what his opponent has in mind. We implement this game in the form of web portal[1] where the user can play that game with the machine. In our game the machine is asking the questions to a human and tries to find what concept the human thought of. The game has been implemented in the limited domain of animals where it shows the abilities to employ the knowledge model that approximates human semantic memory [4] for capturing commonsense relations between concepts [5]. The implementation of elementary linguistic competences employing the cognitive model shows some abilities for human - machine interaction in a way more familiar to humans interactions, but not based on mimicking the language understanding as it is in typical chatterbot conversations.

It should be noticed that the game can be adopted to information retrieval. In typical retrieval systems the user is asking the search engine and as a response he or she obtains the list of results that fulfills the query. In our approach we switch the positions – it is the search engine asking the user questions trying to find out what he or she is searching for.

[1] http://diodor.eti.pg.gda.pl

This reverse in the retrieval process allows to enhance of interaction between human and search engine. The interaction allows to narrow the search results into the set of content relevant to the user requirements.

In our research we implement the presented here algorithm to retrieve the information from Wikipedia. The usage of that repository has several advantages: Wikipedia is a good approximation of WWW nature. On the other hand it is limited and much better structuralized than WWW, thus we find it a very good area to perform the experiments. The structuralization of Wikipedia offers the categories that organize its content. The categories can be used to calculate the queries (used in the word game) asked the user to narrow the search results. In our application, that retrieves the content form Wikipedia, instead of the term `query` (used in the word game) we use notion of `conceptual direction`. The conceptual direction is a phrase calculated by the search engine and presented to the user. The selection of a conceptual direction and assessment of its relevance to the search subject allow to built description of user requirements for retrieval.

2.1 Algorithm

In the algorithm we can select the following steps:

A. Specification of the user's requirements with a search phrase
B. Narrowing the results with the user's interaction
C. Expanding the set of results

In our implementation each of the steps is performed by a separate module thus there can be easily introduced modifications. Additionally to improve effectiveness of the results presentation we introduce their aggregation into groups that fulfill defined conceptual similarity.

Below we describe each of the algorithm steps in details:

A. Keywords Specification

This step is the same as in typical search engines – specification of keywords allows to narrow the set of potential content, relevant to the user's requirements. In our initial implementation we decide to set up granularity level of retrieved information to the set of articles that contains keywords provided by the user. In future we plan to extend it with retrieving particular paragraphs that contain specified keywords. Providing to the user pieces of texts (a bit wider than snippets used in search engines) that contain selected phrases should allow to operate on smaller knowledge chunks and thus be more precise.

It should be noticed that in the algorithm step where keyword-based retrieval is preformed the search phrases can be extended with the usage of additional techniques. Widely used in search engines are methods employing query expansion [6], synonyms analysis, or LSA [7]. They can be easily introduced to our system but as they are not the core of the algorithm currently we use only exact matches between articles content and keywords provided by the user.

We did some research in the area of capturing word co-occurrences employing Hyperspace analog to language (HAL) model [8]. The approach captures statistical co-occurrences of words and thus allows to extend the search phrase and index the content that contains words which are similar (in co-occurring sense) to the search phrases. This functionality will be added soon to the system and we think through the second step of the algorithm the drawback of precision degradation while extending the search phrases will be significantly reduced.

B. Interaction for Narrowing Results

The core element of the algorithm is the interaction with the user that allows to narrow the results according to the user's requirements. In a typical search engine, the requirements can only be defined with search phrases. Here, we also provide the ability for defining the conceptual directions that allows the user to select the way in which he or she wants to continue the search.

In our implementation we employ the fact that Wikipedia articles are organized with categories and the articles with hyper references are related to one another. This allows us to create article representation. For a particular article its computational representation is a combination of article references and the categories they fall into. As both categories and references are associated to the others we were able to enhance representation with higher order associations that have been added with smaller weights.

Employing the associations between categories and articles references allowed us to construct the representation space where the proximity of articles is calculated as the reverse distances between the points. This representation space allow to perform computations on the text. In our approach we calculate the conceptual directions as a hyperplanes that separate the articles in terms of their differences. This difference is captured in the representation space as a category that is the most usable to differentiate articles.

The conceptual direction has been selected as the information gain we obtain if we select particular representation dimension for articles separation.

$$IG_d = -\sum_{i=1}^{I} p_i \log p_i \tag{1}$$

where IG_d denotes information gain calculated for a particular dimension in representation space and p is a normalized value of sums article weights. Formally $p_i = \frac{|w_i|}{|A|}$, where $|A|$ is articles cardinality and $|w_i|$ is a cardinality of particular weight value. The i denotes the number of unique w values in particular dimension d.

It should be noticed the IG measure known from decision trees [9] is calculated for separation of the data set based on cardinality of objects belonging to the same class. Formula 1 performs partitions that maximize division of the dataset into two parts the closest to the half. In the experiments we obtained quite good results with its usage but sometimes if there were significant differences in categories cardinality the information gain performed partitioning of a bigger conceptual set. We plan to reduce this influence, and instead of calculating conceptual directions directly on the articles, calculate them on their prototypes that aggregate similar ones.

The interaction with the user allows to narrow the set of search results to the user's requirements and thus to improve the search precision.

C. Expanding the Result Set

Proximity of the documents can be defined in several ways. The most intuitive and widely used is a number of words co-occurring between them. This value normalized with the number of all words in the article allows to define measure that capture most related articles in terms of their proximity (based on their word content). This approach has been used in popular search engines to introduce retrieval based on similarity but because it has been calculated between single web pages the results were not very satisfying.

In our system we also introduce functionality of extending the selected result set with additional information. The direction of extending the set has been specified by the user that can select particular cluster created within the result set (see section 2.2). The approach we use here calculates the similarity between the group of the most cohesion search results that form a cluster and the new articles from the whole repository that are the most related to that group.

This step requires to calculate article clusters that have been used in presentation layer. The details of the clustering method are described in the next section.

2.2 Presentation Layer

A typical web search engine retrieves the content that contains specified keywords and returns the results sorted with usage ranking function, eg. Google sort its list of the results according to the Page Rank [10]. This approach causes the user typically focuses on the results on the top of the list and leaves other results untouched.

To improve the amount of the data that can be revived by the user we apply content aggregation performed with clustering [11] result set. Based on the representation described in step B of the algorithm we calculate groups of the most similar articles. The user can switch the presentation and see either the ranked list of the results or their groups. It should be noticed that the similarity metrics used to construct groups within search results can be used also to extend this articles set with ones similar but not indexed with keywords (not enclosed in result set).

We examine many approaches to the text clustering and find the most usable here will be the approach based on density analysis that is a modified version of DBSCAN [12]. Despite DBSCAN disadvantage in the computational cost – $O(n^2)$ where n denotes the number of clustered objects the algorithm has many advantages over other approaches. The main one is that it is does not require to explicitly determine the number of clusters, and it can crate clusters having other than convex shapes. The computational resources of our machines allows us to perform clustering process for a thousands of articles in run-time so in our system we apply this approach to organize linear retrieval results into groups. This allows to embrace the search result set more effectively and see its relevant and irrelevant elements.

The selected groups of articles (relevant to user search requirements) can be extended with the approach described in C algorithm step. Based on cluster prototype and similarity measure the result set is extended with additional articles taken form the repository. That offers the additional way for user interaction with information retrieval system and the method for improving the recall of our search engine.

Algorithm 1. Agglomerative DBSCAN algorithm

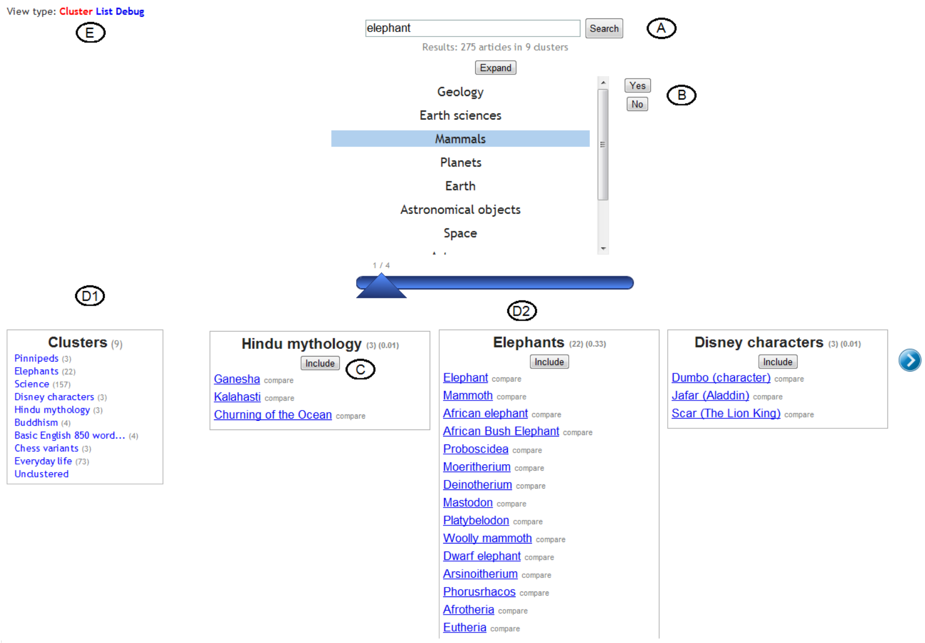
```
$clusters \leftarrow cluster(documents, t_0)$
**for all** $t_i$ in $T$ **do**
    **for all** $c$ in $clusters$ **do**
        **if** $|c| > maxClusterSize$ **then**
            $R \leftarrow R + cluster$
            $delete(cluster)$
        **end if**
    **end for**
    $clusters \leftarrow clusters + cluster(R, t_i)$
**end for**
**for all** $cinclusters$ **do**
    **if** $|c| > maxClusterSize \vee |c| < minClusterSize$ **then**
        $U \leftarrow U + c$
    **else**
        $cat \leftarrow FindRepresentative(c)$
        $docsByCategory[cat] \leftarrow docsByCategory[cat] + c$
    **end if**
**end for**
```

A modification of typical DBSCAN algorithm introduced to our system has been based on performing clustering three times, each time with a different *epsilon* parameter values. The DBSCAN *epsilon* parameter describes the similarity level between two objects to be classified as belonging to one cluster. The succeeding values of *epsilon* parameter we set up empirically to $0.7, 0.6, 0.5$. Three-run DBSCAN allows to combine in each step articles such as at the very beginning are combined together very specific ones and then they form more general groups. The method resembles HAC algorithm [13] with the difference we do not aggregate single articles but the groups of thematically cohesive ones and at the end we do not create hierarchy but flat partitioned clusters. The pseudocode that realizes the above requirements has been presented in algorithm 1.

3 The System Prototype

In Figure 1 we present the system interface that has been deployed as a web page and is available on-line http://bettersearch.eti.pg.gda.pl. In the interface we select the following features:

A. Edit box for entering the search phrase
B. List box for selecting particular conceptual direction and buttons to indicate whether the direction is suitable for the user search or it is not related.
C. Include button that allows to incorporate to the search results additional articles that are related to the specified cluster .
D1. List of the clusters.
D2. Details of the cluster – the cluster label and the articles that belonging to that group.
E. The hyperlinks that allow to switch a type of view: list of results or clusters.


View type: Cluster List Debug
E
elephant
Search
A
Results: 275 articles in 9 clusters
Expand
Yes
No
B
1 / 4
D1
D2
C
Include

Geology
Earth sciences
Mammals
Planets
Earth
Astronomical objects
Space

Clusters (9)
Pinnipeds (3)
Elephants (22)
Science (157)
Disney characters (3)
Hindu mythology (3)
Buddhism (4)
Basic English 850 word... (4)
Chess variants (3)
Everyday life (73)
Unclustered

Hindu mythology (3) (0.01)
Ganesha compare
Kalahasti compare
Churning of the Ocean compare

Elephants (22) (0.33)
Elephant compare
Mammoth compare
African elephant compare
African Bush Elephant compare
Proboscidea compare
Moeritherium compare
Deinotherium compare
Mastodon compare
Platybelodon compare
Woolly mammoth compare
Dwarf elephant compare
Arsinoitherium compare
Phorusrhacos compare
Afrotheria compare
Eutheria compare

Disney characters (3) (0.01)
Dumbo (character) compare
Jafar (Aladdin) compare
Scar (The Lion King) compare

Fig. 1. Web interface of bettersearch system

In the example given in Figure 1 for a specified search phrase *elephant* system retrieves all the articles that contains that word, then they have been organized in clusters. It can be noticed that the articles related to *Religion* have been differentiated with the group of *animals* and *cartoon characters*.

If the user is searching for information about eg. *elephant chess* - *Shogi* he or she using Semantic Sieve algorithm can refine the results. The interaction (succeeding selected conceptual directions) can go as follows:

> search phrase: elephant.
> mammals [NO], Education [No], Religion [No], Art [NO], Board Games [YES]

After finishing selecting the conceptual directions related to the search of Shogi two clusters have been left: *Chess variants* and *Chess*. The user can select one of them and extend the search result set and again refine the results or subsequently review each of them. It should be noticed that if the user searches different information related to the elephant (eg. name of the God Ganeisha) he or she will be selecting different conceptual directions and he would finish with different result set (related to the Hindu Gods).

4 Evaluation

As the implementation of our system is in its initial phase we have defined measures to evaluate its quality. The measure that allows to evaluate the clusters quality we call

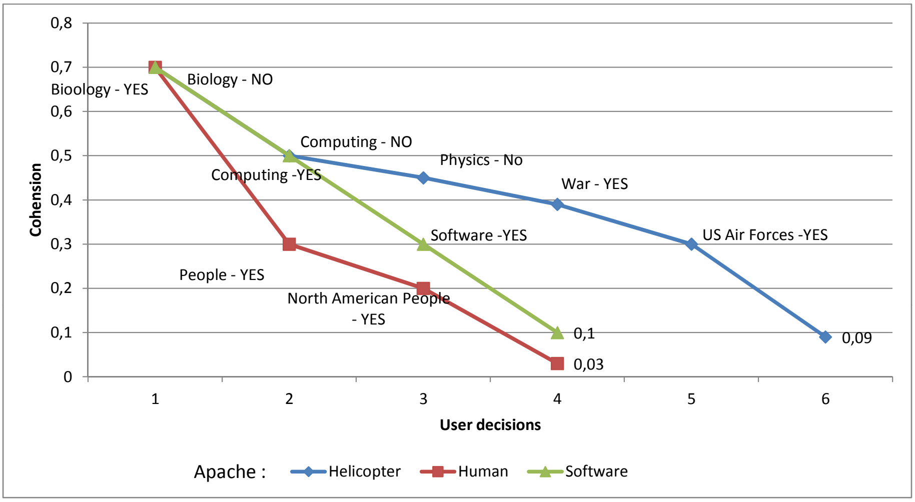


Fig. 2. Convergence of the algorithm for different meaning of keyword Apache

cluster cohesion Ξ_s that is calculated for specified search phrase s. It is computed as a normalized number of irrelevant articles in each of the groups and defined with formula 2 and it is closely related to the well known precision measure.

$$\Xi_s = \frac{\sum_{j=1}^{k} \frac{|a_{j[irr]}|}{|a_j|}}{|a|} \tag{2}$$

where $|a|$ denotes cardinality of all articles $|a_{j[irr]}|$ is cardinality of irrelevant articles in j-th cluster, $|a_j|$ is the j-th cluster cardinality. For a test of 10 search phrases we obtain averaged value $\Xi = 0.781$.

Because of the problems with defining relevance set for the whole repository we were unable to evaluate recall measure exactly. In that case the evaluation of the search process using Semantic Sieve has been performed for 10 multi - sense arbitraryly selected words. For each of them we select particular articles that represent the user's requirements. This constructed relevance set of 37 articles and the evaluation task was to perform the search using conceptual directions according to the user's requirements. The end of the search process has been defined by the user decision and is finished with success if the selected article was in the result set or when the specified article which has been removed from the result set is marked as search failure.

For the test searches 29 have been finished with success. The averaged cohesion of results sets finished with success was $\Xi = 0.89$ which indicates the method keeps stability of the subject that is retrieved. The narrowing of the result set does not influence final results negatively. For the failure searches the averaged cohesion was smaller $\Xi = 0.62$.

The aggregated clusters cohesion for each of the algorithm steps allows us to present conceptual convergence of the algorithm. In Figure 2 we show graphs of convergences for different meanings of the same keyword (here Apache). What can be seen is that the labels of conceptual directions and the user's answers allows to narrow the search results according to the user's preferences despite the fact that we start from the same keyword. Also the set of final results indicates high precision.

5 Future Directions

In the article we present our algorithm called Semantic Sieve for narrowing the search results according to the user's requirements. Initial implementation we made on Polish and Simple Wikipedia. It shows the proposed method can be useful for improving search precision which we plan to evaluate deeper using the proposed methodology on the larger set of test phrases. We also plan to add some improvements to our implementation: - the clusters can be sorted according to relevance to the user's requirements. Improvement of processing effectives will allow us to make computations on English Wikipedia.

To select initial set of articles on which a retrieval with Semantic Sieve is performed we created inverted indexes that allow us to find articles that contain a specified word. As it was mentioned in section 2.1 we plan to extend the exact keyword-match retrieval with methods employing similarity based on on statistical analysis of words co-occurrences in large text collections. As results of evaluation of this approach are very promising [14] we plan to add it to our system to improve the recall of the initial set.

For now we present a list of conceptual directions according to IG score that is computed at the articles granularity level. We think its calculation at the clusters level should considerably improve readability of the directions. We also plan to add some generalizations to conceptual directions which should allow to present more abstract directions that separate the result set better.

In future we plan to extend the evaluation methodology and perform deeper analysis of the results. The main problem here is automatization of the process which each time should be assess by humans. Improving of evaluation methodology will allow to obtain more detailed results that will point to future directions. After successful validation of our search method we plan to extend the repository where retrieval is performed. As a target we want to perform presented search within web pages. To realize this task we are constructing large scale classifier that will allow us to classify web pages into related Wikipedia categories.

Acknowledgements. This work has been supported by the National Center for Research and Development (NCBiR) under research Grant No. SP/I/1/77065/1 SYNAT: "Establishment of the universal, open, hosting and communication, repository platform for network resources of knowledge to be used by science, education and open knowledge society".

References

1. Buckland, M., Gey, F.: The relationship between recall and precision. Journal of the American Society for Information Science 45, 12–19 (1994)
2. Manning, C.D., Raghavan, P., Schütze, H.: Introduction to Information Retrieval. Cambridge University Press, New York (2008)

3. Baeza-Yates, R., Ribeiro-Neto, B., et al.: Modern information retrieval, vol. 463. ACM Press, New York (1999)
4. Quillian, M.: Semantic memory. Semantic Information Processing 2, 227–270 (1968)
5. Szymański, J., Duch, W.: Information retrieval with semantic memory model. Cognitive Systems Research 14, 84–100 (2012)
6. Ogilvie, P., Voorhees, E., Callan, J.: On the number of terms used in automatic query expansion. Information Retrieval 12, 666–679 (2009)
7. Dumais, S.: Latent semantic analysis. Annual Review of Information Science and Technology 38, 188–230 (2004)
8. Lund, K., Burgess, C.: Hyperspace analog to language (hal): A general model of semantic representation. Language and Cognitive Processes (1996)
9. Quinlan, J.: Induction of decision trees. Machine Learning 1, 81–106 (1986)
10. Langville, A., Meyer, C.: Google page rank and beyond. Princeton Univ. Pr. (2006)
11. Carpineto, C., Osiński, S., Romano, G., Weiss, D.: A survey of web clustering engines. ACM Computing Surveys (CSUR) 41, 17 (2009)
12. Ester, M., Kriegel, H.P., Sander, J., Xu, X.: A density-based algorithm for discovering clusters in large spatial databases with noise. In: Proceedings of the 2nd International Conference on Knowledge Discovery and Data mining, vol. 1996, pp. 226–231. AAAI Press (1996)
13. Zhao, Y., Karypis, G., Fayyad, U.: Hierarchical clustering algorithms for document datasets. Data Mining and Knowledge Discovery 10, 141–168 (2005)
14. Szymański, J.: Words Context Analysis for Improvement of Information Retrieval. In: Nguyen, N.-T., Hoang, K., Jędrzejowicz, P. (eds.) ICCCI 2012, Part I. LNCS, vol. 7653, pp. 318–325. Springer, Heidelberg (2012)

GAB-EPA: A GA Based Ensemble Pruning Approach to Tackle Multiclass Imbalanced Problems

Lida Abdi and Sattar Hashemi

Department of Computer Science and Engineering, Shiraz University, Shiraz, Iran
l-abdi@cse.shirazu.ac.ir, s-hashemi@shirazu.ac.ir

Abstract. Processing imbalanced data sets has become a challenging issue in machine learning and data mining communities. Although many researches in the literature have focused on two class problems, multiclass problems have attracted a lot of attention recently. Many existing solutions for multiclass tasks are focused on class decomposition methods, i.e. divide the problem into some two-class sub-problems which are easier to handle.

This paper presents a Genetic Algorithm-Based Ensemble Pruning Algorithm, called GAB-EPA, for multiclass imbalanced problems without applying any class decomposition techniques. In effect, GAB-EPA seeks to find the best subset of classifiers that not only are accurate in their predictions, but also can generate an admissible diversity when gather together as an ensemble model. To show the effectiveness of our approach, we compared our results with other popular ensemble algorithms in terms of three evaluation metrics: Minority Class Recall, G-mean, and MAUC.

Keywords: Diversity, genetic algorithm, ensemble learning, multiclass imbalanced problems, ensemble pruning.

1 Introduction

In recent years, learning from imbalanced data sets has been noticed by industries and academia. In these data sets some classes of data have smaller number of instances compared to other classes. In effect, the objective is to obtain a classifier that will provide high accuracy for the minority class without severely jeopardizing the accuracy of the majority class [6].

Sampling techniques which are comprised of undersampling and oversampling methods, feature selection techniques [8], and ensemble learning algorithms [1,10] have become effective means in processing imbalanced data sets. In particular, ensemble learning algorithms have attracted growing attention in processing imbalanced data due to their effectiveness and proficiency. Bagging, Boosting, mixture of experts and Stacking are among the most famous ensemble learning algorithms that are widely used. For example, SMOTEBoost [1] and RUSBoost [10] are two ensemble methods in processing imbalance data.

A. Selamat et al. (Eds.): ACIIDS 2013, Part I, LNAI 7802, pp. 246–254, 2013.

Even though many efforts so far are concentrated on two class imbalanced problems, multiclass tasks are very common among many real world applications. Protein fold and weld flaw classification are two examples with uneven class distributions [13]. Many existing algorithms which are designed for two class problems, pose many difficulties in multiclass scenarios [11]. Among limited researches in multiclass imbalanced problems, most attention in the literature was dedicated to class decomposition techniques. Class decomposition is referred to as dividing a multiclass problem into a number of two class sub-problems. OAA and OAO are two most popular techniques of class decomposition in the literature.

In this paper we propose a Genetic Algorithm-Based Ensemble Pruning Algorithm, called GAB-EPA to deal with multiclass imbalanced data sets without applying any class decomposition schemes. In the first step of training, classifiers are generated as the initial search space of the GA algorithm. After applying genetic algorithm and finding the best subset of classifiers, a weighting procedure is applied to assign a weight to each ensemble member. The final output of the test sample would be the weighted vote of the ensemble members. Our results showed that our method can compete against popular ensemble learning methods, Bagging and AdaBoost, and even outperform them in most cases.

The rest of the paper is organized as follows. In section 2 some related researches will be discussed. Section 3 will explain our proposed method, GAB-EPA, in more details. Our experimental results are given in Section 4. Section 5 concludes the paper by a conclusion part and presents the future work.

2 Related Work

Sun et al. [11] integrated a cost-sensitive boosting algorithm with genetic algorithm. They used genetic algorithm to search the optimum cost setup of each class and develop a cost-sensitive ensemble algorithm, named AdaC2.M1, to tackle multiclass problems.

Wang and Yao [13] incorporated negative correlation learning, with AdaBoost [4]. A new negative correlation-based algorithm, called AdaBoost.NC, handles multiclass imbalanced problems. They adapt the ambiguity term, which is a measure of diversity in regression tasks, for classification problems.

We used these ideas and designed GAB-EPA. In general, the ensemble learning process is comprised of three steps. The first step is ensemble generation, which is consists of generating a set of models. Usually, during generation, some of the models are redundant. In the second step, the ensemble pruning step, the ensemble is pruned by eliminating some of the models. Finally, in the ensemble integration step, a strategy to combine the base models is defined. This strategy is then used to obtain the prediction of the ensemble for test instances, based on the predictions of the base individuals [9].

3 The Proposed Method: GAB-EPA

In this section GAB-EPA algorithm will be discussed. At the first step of ensemble learning, n number of base learners will be trained. During genetic algorithm, $m(m < n)$ number of these learners which are accurate and can yield acceptable diversity together will be selected. In integration part of ensemble learning, these selected learners will get a weight, according to their G-mean values. In test phase, the weighted vote of these classifiers will be the output of the coming case. The proposed method GAB-EPA is explained in Algorithm 1.

Algorithm 1. The Procedure of the GAB-EPA Algorithm

```
Input: Train Data $S : (x_1, y_1), ..., (x_i, y_i), (x_t, y_t)$, GAB-EPA parameters,
class labels $y \epsilon \{1, 2, ..., c\}$ .
Output: Final ensemble model


Step 1. Select n sets of training data with t samples in each
        from S, using sampling with replacement.
Step 2. Train n classifiers on n sets of training data.
Step 3. Begin Genetic Algorithm.
        Build a random initial population of size N,each
        chromosome is of length m.
FOR i=1 to g
        -Do parent selection,crossover and mutation considering
         crossover and mutation rates;
        -Generate N new offsprings and add them to the population;
    -Compute fitness of all individuals in the population;
    -Transfer N best individuals to the next generation;
END FOR

Step 4. Select the best individual of all generations and create
        the ensemble.
Step 5. Assign a weight to each classifier in ensemble according
        to its G-mean value, then normalize the weights:
```

$$W_i = \frac{G-mean_i}{\sum_{i=1}^{m} G-mean_i} .$$

(For each coming test instance, the output will be the weighted vote of ensemble members.)

3.1 Generation

In generation part, n total number of base learners will be trained. We used $C4.5$ decision tree as our base classifier, which is implemented in Weka [14], with the default parameters.

The training sets of data which are used to train each of these n classifiers on, should be different from each other. If all classifiers train on the same sets of data, then all the classifiers will be identical and diversity among them would not be significant. Therefore we used sampling with replacement technique in order to lead the classifiers to be trained on different aspects of data. The block diagram of the GAB-EPA is given in figure 1.

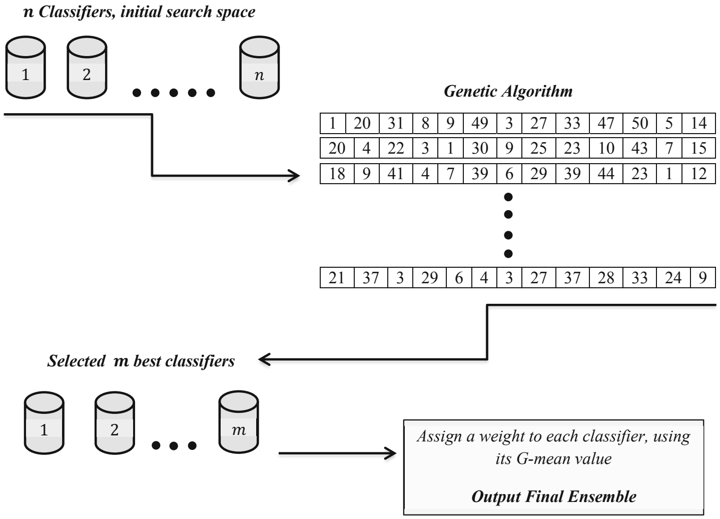


Fig. 1. GAB-EPA uses n classifiers as its initial search space. It selects m classifiers with the aid of genetic algorithm. A set of learners that are both accurate and diverse is used as the final ensemble model.

3.2 Individual Representation, Fitness Function and Genetic Operators

Each individual in GAB-EPA is a vector of length m, corresponding to m indexes of classifiers. We set m to 13 in all algorithms. Classifiers' indexes vary from 1 to n, in which we set n to 50. So, the initial population of the GAB-EPA will be comprised of N number of such individuals. The initial population size is set to 100 and we choose the best individual of all 50 generations.

In order to fulfil the need of accurate and diverse classifiers, the fitness function should be designed in such a way that the final model could best fit the multiclass imbalance problem. Consequently we define the fitness of each population member i as follows:

$$Fitness_i = \frac{1}{m}\sum_{k=1}^{m} G-mean_k + diversity_{(i)} \quad . \tag{1}$$

For two-class data sets, $T = \{t_1, t_2, ..., t_N\}$ is composed of (x_i, y_i) where y_i is the true label of instance x_i and $y_i \in \{1, 0\}$. For each classifier $f_i(i = 1, ..., L)$ $y_i = 1$ if f_i classifies t_i correctly and 0 otherwise. The extended G-mean [11] for each classifier, which is the geometric mean of the recall of all classes, is calculated as follows:

$$G-mean = (\prod_{i=1}^{m} R_i)^{\frac{1}{c}} \quad . \tag{2}$$

$$R_i = \frac{TP_i}{TP_i + FN_i}, i = 1, ..., c \ . \tag{3}$$

Where c is the number of classes, in multiclass case ($c > 2$). TP is the number of instances of class i which are classified correctly and FN is the number of instances that are misclassified.

In order to calculate the individual's fitness, we used whole training data. So we calculate recall of each class separately, considering all instances of one class as positive and calculate the recall using equation (3) and then take the geometric mean to compute extended G-mean. Since we have m indexes of classifiers in each chromosome, we can get the average of calculated G-mean over all classifiers.

In order to calculate the diversity, we used Q-statistic [7] which is a pairwise measure of diversity and is particularly recommended for its simplicity and understandability [7,12]. The Q-statistic will be calculated for two classifiersf_i, f_k as follows:

$$Q_{i,k} = \frac{N^{11}N^{00} - N^{01}N^{10}}{N^{11}N^{00} - N^{01}N^{10}} \ . \tag{4}$$

Where N^{ab} is the number of t_j of T for which $y(i, j) = a$ and $y(j, k) = b$. Q is the average over all pairs of classifiers. For example we have m classifiers in the given individual, then we have $\frac{m(m-1)}{2}$ pairs of classifiers. Higher Q-statistic indicated smaller diversity and lower disagreement between classifiers. As a result *diversity* =1-(Q-statistic)$\in [-1, 1]$. In each generation, the better the chromosome the higher the fitness.

Crossover takes two chromosomes as inputs which were chosen through tournament parent selection of size three. Crossover is one point. In other words, a breakpoint in range $[0, m-1]$ is chosen randomly and then creating two children by exchanging the tails [2].

In mutation, if a gene of an individual is selected for mutation, its value will substitute for the randomly chosen index among all indexes that are not contained in the chromosome. It worth mentioning that, during crossover and mutation, no individual will contain the repeated index. We set the mutation and crossover rates to 0.05 and 0.85, respectively.

4 Experimental Settings

4.1 Data Sets and Performance Metrics

The performance of our ensemble learning algorithm is evaluated using ten benchmark imbalanced data sets of UCI [3]. Table 1 provides some characteristics of these data sets. In each one, we used twenty percent as test set and the rest as train set. According to the class distributions which are given in Table 1, we have three kinds of data sets.

We used three different performance metrics to evaluate the performance of implemented algorithms. In each data set, we consider one class as positive class

Table 1. Description of benchmark data sets. There are three kinds of data sets: 1. Multi-minority, 2. Multi-majority and 3. Multi-minority/Multi-majority.

Data set	Class	Size	Distribution	Type
New-Thyroid	3	215	150/35/30	1
Ecoli	5	327	143/77/52/35/20	1
Cleveland	5	303	164/55/36/35/13	1
Page	5	5473	4913/329/28/88/115	1
Wine	3	178	48/59/71	2
Contraceptive	3	1473	629/333/511	2
Balance	3	625	49/288/288	2
Solarflare2	6	1066	147/211/239/95/43/331	3
Satimage	6	6435	1533/703/1358/626/707/1508	3
Dermatology	6	366	112/72/61/52/49/20	3

and calculate the corresponding metric. We can compute G-mean and minority class recall using equation (2) and (3) respectively.

MAUC, the extended AUC in multiclass cases is the average of AUC over each pairs of classes. We can compute MAUC as follows:

$$MAUC = \frac{2}{c(c-1)} \sum_{i<j} \frac{[A(i,j) + A(j,i)]}{2} \ . \tag{5}$$

Where $A(i,j)$ is the AUC between class i and class j calculated from the ith column of M. M is the $t \times c$ matrix which is provided by classifier. Each element of M, m_{pq} indicates the probability of belongingness of instance p to class q. We should be aware that in multiclass cases $A(i,j), A(j,i)$ may not be equal, so both of them should be taken into account.

4.2 Computational Overhead of the GAB-EPA Algorithm

Some of the factors that affect the time complexity of the genetic algorithm are representation of the individuals, length of each chromosome, population size, genetic operators, and the fitness function. We used integer representation. The initial population size is set to 100 and the length of each chromosome is set to 13. The genetic operators' complexity overheads which are tournament parent selection, one point crossover, and per-bit mutation can be computed in polynomial time.

The most important thing is the fitness evaluation. In GAB-EPA, each of the initial classifiers classifies the train examples completely before the genetic algorithm begins. The second term of the fitness function calculates the diversity which can be computed in polynomial time. Given all these factors, the overall complexity overhead of the genetic algorithm is polynomial.

4.3 Analyses and Observations

The performance of the proposed method and two other ensemble learning algorithms in terms of minority class recall, G-mean, and MAUC is evaluated over ten UCI datasets. The average of ten runs over each data set is reported. The means and standard deviations these metrics are computed. The results are summarized in Table 2. Recall is a measure of completeness, it indicates that how many positive class instances are labelled correctly. Extended G-mean, on the other hand, indicates that how well a classifier can balance the recognition among different classes [13]. MAUC assesses the average ability of separating any pair of classes from each other. In this aspect and in comparison to most data sets our results showed that our method outperforms the classic ensemble learning algorithms in terms of minority class recall, G-mean, and MAUC.

Not surprisingly, our method performs better in recalling minority class instances, because the algorithm is converged to generate an accurate and diverse ensemble model. As it was shown in [12], diversity shows a positive impact on the minority class in general; it is also beneficial in the overall performance in terms of AUC and G-mean. Higher G-mean indicates that the algorithm does not ignore any class of data and can perform with equal respect to all classes especially minority class. Due to the investigations that are conducted by [12], increasing diversity in ensembles is beneficial in terms of AUC and MAUC. Our results in three of the data sets i.e., *Solarflare2*, *Balance* and *Dermatology* are not as good as others. This is probably related to the type of data sets, multi-minority/multi-majority, and multi-majority. It seems that GAB-EPA has difficulty in dealing with these kinds of data sets. Being diverse and accurate, members of ensemble are a key factor in ensemble learning.

5 Conclusion and Future Work

In this paper we proposed a GA-based ensemble pruning algorithm, GAB-EPA, which employs genetic algorithm in order to find a diverse and accurate subset of classifiers. It is really important that these two factors are used together. In order to do so, we designed a fitness function which measures the diversity and G-mean of the population members and finds the best ensemble model. Our results over ten benchmark data sets showed promising results, in comparison with two of the most popular ensemble learning algorithms.

Future work will continue to investigate the performance of GAB-EPA with different base classifiers. Different stable and unstable base learners may affect the performance of the algorithm; also using other diversity measures may have good results in our algorithm.

Table 2. Means and Standard Deviations of Minority Class Recall,G-mean and MAUC by Bagging, AbdBoost and GAB-EPA over 10 datasets of UCI. Minority Class Recall is computed for the Smallest Class of each dataset. Greater values are highlighted in boldface.

Minority-Class Recall			
	Bagging	AdaBoost	GAB-EPA
New-Thyroid	0.883 ± 0.080	0.800 ± 0.204	**0.966 ± 0.070**
Ecoli	0.70 ± 0.229	**0.825 ± 0.12**	0.825 ± 0.205
Cleveland	0.033 ± 0.105	0.033 ± 0.105	**0.066 ± 0.14**
Page	0.816 ± 0.183	0.766 ± 0.195	**0.833 ± 0.136**
Wine	0.94 ± 0.084	0.97 ± 0.048	**0.980 ± 0.042**
Contraceptive	0.3611 ±0.065	0.353±0.066	**0.368±0.056**
Balance	0.020 ± 0.042	**0.06 ± 0.084**	0.0 ± 0.0
Solarflare2	0.1 ± 0.110	**0.266 ± 0.140**	0.166 ± 0.130
Satimage	0.562 ± 0.036	0.569 ± 0.034	**0.597 ± 0.031**
Dermatology	0.975 ± 0.079	**1.0 ±0.00**	0.800± 0.283
G-mean			
	Bagging	AdaBoost	GAB-EPA
New-Thyroid	0.904±0.062	0.880±0.079	**0.947±0.047**
Ecoli	0.756±0.064	0.726±0.052	**0.805±0.047**
Cleveland	0.027 ±0.086	0.030 ±0.097	**0.060 ± 0.128**
Page	0.833 ±0.046	0.750 ±0.054	**0.846 ± 0.029**
Wine	0.934 ±0.040	0.943 ±0.040	**0.977 ± 0.027**
Contraceptive	0.486 ± 0.038	0.472 ± 0.031	**0.487± 0.029**
Balance	0.085 ± 0.18	**0.185 ± 0.241**	0.0 ± 0.0
Solarflare2	0.304 ±0.265	**0.501 ±0.054**	0.360 ± 0.253
Satimage	0.853 ± 0.011	0.847 ± 0.011	**0.865 ± 0.009**
Dermatology	**0.976 ± 0.023**	0.969 ±0.021	0.930 ± 0.0705
MAUC			
	Bagging	AdaBoost	GAB-EPA
New-Thyroid	0.979±0.037	0.991±0.008	**0.997±0.003**
Ecoli	0.938±0.026	0.935±0.013	**0.951±0.029**
Cleveland	0.645 ±0.043	0.635±0.042	**0.660 ±0.057**
Page	0.972 ±0.008	0.967±0.014	**0.976 ±0.009**
Wine	0.991 ±0.008	0.995±0.005	**0.998 ±0.001**
Contraceptive	**0.704 ± 0.026**	0.687±0.018	0.696±0.0196
Balance	0.796 ± 0.032	**0.817 ± 0.029**	0.802 ± 0.023
Solarflare2	0.845 ±0.019	0.850±0.023	**0.860±0.015**
Satimage	0.979 ± 0.002	0.977± 0.002	**0.981± 0.002**
Dermatology	0.992±0.006	**0.997 ±0.002**	0.995±0.005

References

1. Chawla, N.V., Lazarevic, A., Hall, L.O., Bowyer, K.W.: SMOTEBoost: Improving Prediction of the Minority Class in Boosting. In: Lavrač, N., Gamberger, D., Todorovski, L., Blockeel, H. (eds.) PKDD 2003. LNCS (LNAI), vol. 2838, pp. 107–119. Springer, Heidelberg (2003)
2. Eiben, A.E., Smith, J.E.: Introduction to Evolutionary Computing, 1st edn. (2003)
3. Frank, A., Asuncion, A.: UCI machine learning repository, `http://archive.ics.uci.edu/ml`
4. Freund, Y., Schapire, R.E.: A Decision-theoretic Generalization of Online Learning and An Application to Boosting. In: Vitányi, P.M.B. (ed.) EuroCOLT 1995. LNCS, vol. 904, pp. 23–37. Springer, Heidelberg (1995)
5. Hand, D.J., Till, R.J.: A simple generalization of the area under the ROC curve for multiple class classification problems. Machine Learning 45(2), 171–186 (2001)
6. He, H., Garcia, E.A.: Learning from imbalanced data. IEEE Transactions on Knowledge and Data Engineering 21(9), 1263–1284 (2009)
7. Kuncheva, L.I., Whitaker, C.J.: Ten measure of diversity in classifier ensembles: Limits for two classifiers. In: A DERA/IEE Workshop on Intelligent Sensor Processing (Ref. no. 2001/050), pp. 10/1-1010. IET (2001)
8. Alibeigi, M., Hashemi, S., Hamzeh, A.: DBFS: An effective density based feature selection scheme for small sample size and high dimensional imbalanced data sets. Data and Knowledge Engineering 8182, 67–103 (2012)
9. Mendes-Moreira, J., Soares, C., Jorge, A.M., Sousa, J.F.D.: Ensemble approaches for regression: A survey. ACM Computing Surveys (CSUR) 45(1), 10 (2012)
10. Seiffert, C., Khoshgoftaar, T., Hulse, J.V., Napolitano, A.: Rusboost: A hybrid approach to alleviating class imbalance. IEEE Transactions On Systems, Man, Cybernetics-Part A 40(1), 185–197 (2010)
11. Sun, Y., Kamel, M.S., Wang, Y.: Boosting for learning multiple classes with imbalanced class distribution. In: ICDM 2006: Proceedings of the Sixth International Conference on Data Mining, pp. 592–602 (2006)
12. Wang, S., Yao, X.: Relationships between diversity of classification ensembles and single-class performance measures. IEEE Transactions on Knowledge and Data Engineering 25(1), 206–219 (2011)
13. Wang, S., Yao, X.: Multiclass imbalance problems: Analysis and potential solutions. IEEE Transactions On Systems, Man, Cybernetics-Part B 42(4), 1119–1130 (2012)
14. Witten, I.H., Frank, E.: Data Mining: Practical Machine Learning Tools and Techniques, 2nd edn. Morgan Kaufmann, San Francisco (2005)
15. Zhao, X.M., Li, X., Chen, L., Aihara, K.: Protein classification with imbalanced data. Proteins: Structure, Function, and Bioinformatics 70, 1125–1132 (2008)

Performance of Different Techniques Applied in Genetic Algorithm towards Benchmark Functions

Seng Poh Lim and Habibollah Haron

Department of Computer Science, Faculty of Computing,
Universiti Teknologi Malaysia,
81310 Skudai, Johor, Malaysia
lawrencess87@yahoo.com,
habib@utm.my

Abstract. Optimisation is the most interesting problems to be tested by using Artificial Intelligence (AI) methods because different optimal results will be obtained when different methods are implemented. Yet, there is no exact solution from the methods implemented because random function is usually applied. Genetic algorithm is a popular method which is used to solve the optimisation problems. However, no any methods can execute perfectly because the way of the method performs is different. Therefore, this paper proposed to compare the performance of GA with different operation techniques by using the benchmark functions. This can prove that different techniques applied in the operations can let GA produces different result. Based on the experiment result, GA is proved to perform well in the optimisation problems but it highly depends on the techniques implemented. The techniques for each operation have shown different performance in obtaining the time, minimum and average values for benchmark functions.

Keywords: Genetic Algorithm, Optimisation, Benchmark Functions, Performance.

1 Introduction

Optimisation is the most interesting problems to be tested by using Artificial Intelligence (AI) or Soft Computing methods. This is because different optimal results will be obtained when different methods are implemented. However, there is no exact solution from the methods implemented because random function is usually applied in AI method. Basically, same optimisation problems were tested by using different methods because the researchers want to test the performance of the method in obtaining the result. If the result obtained is the most minimum and approximate to the actual optimum solution, hence this method is considered contains the best performance compared to the others.

Genetic algorithm (GA) is a popular soft computing technique which is used to solve the optimisation problems. Besides that, it can also be used to deal with scheduling, surface fitting and searching problems. GA has been widely used to deal with different kind of optimisation problems and it able to find and produce a good

A. Selamat et al. (Eds.): ACIIDS 2013, Part I, LNAI 7802, pp. 255–264, 2013.

solution. Some of the previous works also compared the performance of GA with other methods. However, different methods contain its pros and cons while using it to deal with the problems.

Some methods maybe performed well and some maybe not good. Hence, this can be concluded that no any methods can execute perfectly because the way of the method performs is different. Some methods maybe require more iteration for it to get better solution while some maybe due to the parameter setting problems. Same goes to GA, the techniques in the operations maybe is one of the problems which affect the performance of GA in obtaining the optimum result. Therefore, this paper proposed to compare the performance of GA with different operation techniques by using the benchmark functions. This can prove that although this is the same GA method, but different techniques applied in the operations can produce different result. There are a lot of different methods in the GA operations which can let GA generate different result for the same problems. Hence, the performance of GA is a good research field to be further explored.

This paper is structured as follows. In Section 2, we review the theories and previous works on Genetic Algorithm. Section 3 describes flow of experiment in testing the benchmark functions using 2 GA models. Analysis and discussion on the experimental results are demonstrated in Section 4. Conclusions and future work are presented in the last section.

2 Review on Genetic Algorithm

Genetic Algorithm (GA) method was introduced by John Holland in 1975 and he developed this idea in “Adaptation in natural and artificial systems” [1]. GA is suitable to be used as a tool in solving the searching and optimisation problems. It is based on the principle of genetics and also evolution, which is similar to other soft computing methods [2] such as Evolution Strategies and Genetic Programming.

GA contains several operations in performing the algorithm. In each of the operation, it contains several techniques. The techniques used will produce different result because the way of techniques perform are different. Indirectly, the performance of GA in obtaining the result will be affected by the technique used. As stated by [3], suitable technique will increase the performance of GA. Better result is important especially in the optimisation problem since optimal result is preferred.

A set of population is generated in order to perform the GA operations. Generally, single population is used in testing the problem. However, GA can contain more than 1 population, which is known as Island Genetic Algorithm [4]. Each population or known as island can contain subpopulation. GA operations are occurred independently from other subpopulations and the individuals can migrate to other population [4]. For each population, it contains several individuals or refers as chromosomes. These chromosomes will be sub divided into several genes [2]. Normally, random function is used to generate the genes for each chromosome.

The chromosomes in the population will be evaluated based on the objective or fitness function. Fitness function is derived from objective function [5] and is used to evaluate the strength of the chromosome toward the case study [4,6]. As stated by [4], fitness function is the major component in Evolutionary Algorithm. It demonstrates how well the solution and close toward the optimal value [1]. This is because for optimisation problem, it must contain all the criteria to be optimised [4]. The fitness values will show how good the solution by based on the minimising or maximising problem [2]. After this process, it proceeds with the selection, crossover and mutation operation in producing the next generation chromosomes.

Most of the selection techniques will choose the parent based on their fitness value. The chromosome that contains higher fitness will be selected more frequently compared to others in producing the children. This is because the chromosome that contains good genetic materials which can obtain better result should be inherited to the next generation. The selection method will affect the convergence speed and result obtained. Hence, it is the most important step.

While for crossover and mutation operations, it is highly based on the probability generated by the algorithm [2]. If the generated probabilities are less than compared to the default probabilities, hence both of the operations will be carried out or vice versa [2]. Basically, there are a lot of crossover and mutation operations techniques and they will introduce new genetic materials to the population. So, it will maintain the diversity of the population [1] and improve the search space in searching optimum solution. For crossover, the genetic materials of 2 parents will be combined and new generation children will be produced. While for mutation, it will insert some new or change a bit the genetic materials of the chromosome. Better solution can be produced through these operations. However, sometimes these operations will lead to even worse result.

Convergence criteria are used to determine when the operations will stop in producing the next generation children. Most of the research work will use maximum generation as the convergence criteria. Hence, new offsprings (children) will be produced after all the operations were finished. All the new offsprings will be evaluated using the same objective or fitness function [2]. The last step will be the replacement operation. If the new offsprings (children) are better compared to the old generation (parents) based on the fitness values, hence the old generation chromosomes will be replaced with the new generation chromosomes [2].

Several previous researches and related works were studied for this method. GA is applied in [7] work to reconstruct the shape and the dielectric constant of the object. The cost function is minimised by using GA in the inversion procedure. While in [8] work, GA is used because it contains the characteristics such as random, iteration and evolution, hence it is suitable to be applied in managing the data problems. As for [9] work, GA is applied to solve the optimising process in the inverse problems by changing the parameters in the next iterations with some rules. GA is implemented in [10] work to deal with the minimisation problems on difference between the real and predicted measurement. The parameters are coded into GA and recovered through the global optimisation.

Most of the papers focused on the selection techniques because the selection operation will determine which parent will be selected in producing the children for the next generation. In [11], the selection techniques in GA have been compared extensively. All selection techniques will select the best fitness parent as the parent in producing the children as he stated. However, there are pros and cons among the selection techniques. For [12], they make a comparison among the selection techniques (proportionate, ranking, tournament and genitor). Based on them, ranking and tournament are performed better compared to proportionate technique. While [13] stated that elitism is the best selection technique compared to Roulette wheel and tournament by testing the selection techniques on TSP. In [14], two crossover operators are applied in solving the job shop scheduling problems and good result is obtained. This implies that hybrid method can produce better result. Steady state and generational update are compared in [3] work by testing on nonstationary environment case study. Steady state update is proved to be better compared to generational update because it will straightly use the updated chromosome on the mating pool, hence better result will be produced.

3 Flow of Experiment

This paper makes a comparison between two GA models of different operation techniques by obtaining minimum value for all benchmark functions. The model that can produce the smallest value and approximate to the actual optimum is considered as the best model with good performance. Same flowchart as shown in Fig. 1 will be used by both GA models but with different operation techniques.

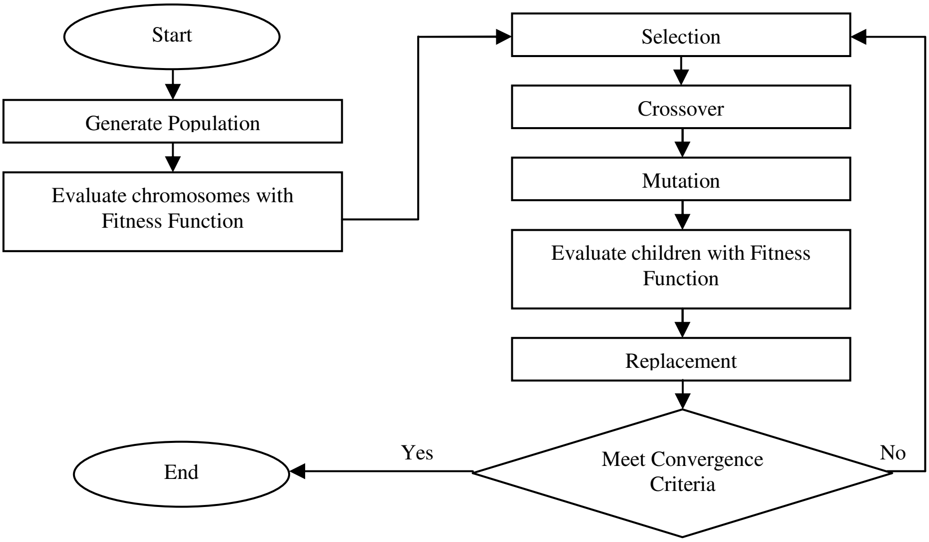

Start
Generate Population
Evaluate chromosomes with Fitness Function
Selection
Crossover
Mutation
Evaluate children with Fitness Function
Replacement
Meet Convergence Criteria
Yes
No
End

Fig. 1. Flowchart of Genetic Algorithm

The model is performed as below:

1. A set of population is randomly generated in the range given based on the benchmark function.
2. The chromosomes from the population are evaluated using the fitness function.
3. Selection: 2 set of chromosomes are selected to be the parent in producing the children.
4. Crossover: 1 random number is generated. If this random number is less than the crossover probability, hence crossover operation is performed. Else, continue to the next operation.
5. Mutation: 1 random number is generated. If this random number is less than the mutation probability, hence mutation operation is performed. Else, continue to the next operation.
6. The children are evaluated using the fitness function.
7. Replacement: Old generation parents are replaced by new generation children.
8. If termination criteria are not fulfilled, continue to 3. Else, the algorithm is terminated.

Table 1 shows the different operation techniques applied in GA1 and GA2 model. For GA1 model, random selection is used where the parents are selected randomly in producing the children. Single point crossover is used by randomly selecting one point from the chromosome and break on the point. The head of chromosome for parent 1 and 2 before the break point are combined with the tail of chromosome for parent 2 and 1 in producing the children. Interchanging mutation is applied by randomly selecting 2 points from the chromosome and swapping the value of the points. Steady state replacement is applied by replacing the weak parents with the stronger children for each generation.

For GA2 model, tournament selection is applied by randomly selecting 4 chromosomes from the population. The 2 chromosomes with the highest fitness value among the 4 chromosomes will be selected as the parent. Uniform crossover is used by generating 1 crossover mask with the bit value of 0 and 1. The crossover is performed based on the mask value. If the value of mask is 1, the genes from parent 1 are directly copied to child 1 and parent 2 to child 2. But if the value of mask is 0, crossover is performed by copying the genes from parent 1 to child 2 and parent 2 to child 1. For random mutation, 2 points are randomly selected from the chromosome and the values are replaced with a random value generated in the fixed range. Generation update is applied by replacing all the old population chromosomes with the new population chromosome. None of the previous generation chromosomes will exist in the new population. Maximum generation is used as the convergence criteria to terminate the algorithm for both models.

Table 1. GA models with different operation techniques

No.	Operation	GA 1	GA 2
1.	Selection	Random	Tournament
2.	Crossover	Single point	Uniform
3.	Mutation	Interchanging	Random
4.	Replacement	Steady state update: Weak parent replacement	Generation update

For this experiment, same parameters setting are used for both GA models. Same setting is applied because this just can show and prove the performance of both models in obtaining minimum value. Besides that, the parameters setting are depending on user defined. Each benchmark function will contain different dataset which is generated randomly in the "Generate Population" step as shown in Figure 1. There are no any standard dataset in testing the benchmark functions. This experiment is conducted by using Dev C++. Table 2 shows the parameters setting for both GA models.

Table 2. Parameters setting for both GA models

No.	Parameter	Value
1.	Number of generation	2000
2.	Population size	40
3.	Number of dimension	30
4.	Crossover probability	0.7
5.	Mutation probability	0.01
6.	Number of testing	10

Equations 1 – 5 are the benchmark mathematical functions with the range value of dataset which are used to test the performance of GA. The benchmark functions are referred from [15-17] and the global optimum value for all these functions are 0. Please refer to [15-17] for the details of each benchmark function. These benchmark functions are used as the fitness function in GA. Minimum value which approximates to the actual optimum is considered as the result for each function, which are also the optimised criteria for this experiment.

1. Sphere function

$$f_1(x) = \sum_{i=1}^{n} x_i^2 \tag{1}$$

$-5.12 \le x_i \le 5.12, i = 1,..., n$
Global minimum, $f_1(x) = 0$ for $x_i = 0, i = 1,..., n$
n is the number of dimension

2. Ackley function

$$f_2(x) = 20 + \exp(1) - 20\exp\left(-0.2\sqrt{\frac{1}{n}\sum_{i=1}^{n} x_i^2}\right) - \exp\left(\frac{1}{n}\sum_{i=1}^{n}\cos(2\pi x_i)\right) \tag{2}$$

$-30 \le x_i \le 30, i = 1,..., n$
Global minimum, $f_2(x) = 0$ for $x_i = 0, i = 1,..., n$
n is the number of dimension

3. Rastrigin function

$$f_3(x) = 10n + \sum_{i=1}^{n} (x_i^2 - 10\cos(2\pi x_i)) \tag{3}$$

$-5.12 \le x_i \le 5.12, i = 1,..., n$

Global minimum, $f_3(x) = 0$ for $x_i = 0, i = 1,..., n$

n is the number of dimension

4. Zakharov function

$$f_4(x) = \sum_{i=1}^{n} x_i^2 + \left(\sum_{i=1}^{n} 0.5x_i\right)^2 + \left(\sum_{i=1}^{n} 0.5x_i\right)^4 \tag{4}$$

$-5 \le x_i \le 10, i = 1,..., n$

Global minimum, $f_4(x) = 0$ for $x_i = 0, i = 1,..., n$

n is the number of dimension

5. Axis parallel hyper-ellipsoid function

$$f_5(x) = \sum_{i=1}^{n} ix_i^2 \tag{5}$$

$-5.12 \le x_i \le 5.12, i = 1,..., n$

Global minimum, $f_5(x) = 0$ for $x_i = 0, i = 1,..., n$

n is the number of dimension

4 Experimental Results

As stated by [5], there is no mathematical proof in testing the convergence of GA. No exact solution will be obtained from the GA method. Hence, this experiment is only to test the performance of 2 GA models by obtaining minimum value which approximate to the actual optimum solution and also time taken in producing the result.

Table 3 shows the experiment result produced by using GA1 and GA2 model. Each benchmark function is tested 10 times for both models. The minimum and maximum optimum values among 10 times testing are recorded. This is to show how much optimum value can be produced by using the same model. Then, average of 10 times testing for all functions are recorded and considered as the performance measurement. Besides that, average time in producing the result has been taken as another comparison for both models.

Based on result shown in Table 3, it implies that GA1 model performs better compared to GA2 in obtaining minimum value. Besides that, less time was used in producing the result for GA1 model. Both models show a huge difference in obtaining the average results for all functions although the same parameters setting were used in testing the performance. The result also shows that the techniques used for each operation in GA are important because it will affect the time taken and result produced.

Table 3. Experiment result

No	Function	Actual Optimum	Model	Min	Max	Average	Average Time (s)
1	Sphere	0	GA1	0.000	0.198	0.039	0.064
			GA2	19.076	22.490	20.873	0.408
2	Ackley	0	GA1	0.017	0.433	0.187	0.077
			GA2	11.520	12.726	11.987	0.652
3	Rastrigin	0	GA1	0.000	0.381	0.128	0.076
			GA2	102.688	118.223	112.725	0.642
4	Zakharov	0	GA1	0.082	0.507	0.280	0.065
			GA2	132.570	167.170	157.012	0.437
5	Axis Parallel Hyper - Ellipsoid	0	GA1	0.002	0.748	0.298	0.063
			GA2	175.773	277.674	258.424	0.425

GA1 model shows that it able to get the actual optimum result for Sphere and Rastrigin function although random function is performed in selecting the parent. Besides that, it also shows good performance in obtaining the minimum value which approximate to the actual optimum value for Ackley, Zakharov and Axis parallel hyper-ellipsoid function. As for average, GA1 model shows that it able to get the average less than 0.3 compared to the actual optimum for all functions. GA1 model shows that it performs faster than GA2 model because the average time taken for GA1 model in producing the result is less than GA2 model for all functions.

While for GA2 model, it shows bad performance in obtaining minimum and average values for all functions. Furthermore, more time was taken in producing the result compared to GA1 model. This model obtains poor result maybe is due to the techniques applied are not suitable and more iteration is needed for it to perform better. In addition, the parameters involved might affect the result obtained for both models. If the parameters are assigned with higher value, convergence toward the optimum value maybe is faster because crossover and mutation operations will keep on performing, but maybe it will trap in local minima. Although this will improve the searching space, but crossover and mutation operations will destroy the structure of the best chromosome. Hence, more iteration is needed in searching better result. The characteristics of techniques applied in GA1 and GA2 model which have affected the result are discussed as below.

For selection operation, although tournament technique has included the fitness value in selecting the parent, but not all the best parents are selected in producing the children. Sometimes best parent maybe will produce even worse children. But for random technique, it just randomly selects the parent from the population and produces the children. It does not depending on the fitness of the parent. Sometimes bad parents can produce good children. Maybe this is the reason which let the random technique performs better and obtains good result compared to tournament technique.

While for crossover, uniform crossover maybe is better compared to single point crossover because mask chromosome is used in performing the crossover operation. Hence, it will improve the search space in searching for optimum value.

However, sometimes the structure of best parent is destroyed due to the crossover only take place based on the mask generated in producing the children. While for single point crossover, the head of chromosome for parent 1 is joined with the tail of parent 2. So, the children produced can still maintain the structure of best parents and maybe increase the overall fitness value of children.

Random mutation maybe even better because the random function is applied by introducing the new genetic materials to the chromosome and increase the search space. But, the flipping might introduce even worse value to the chromosome. Interchanging maybe is less effective if compared to random mutation because it only swaps the 2 points value and does not introduce any new genetic materials. However, this technique will apply its effect in the next generation because crossover will take place and cross the gene. Hence, better result can be produced.

Generation update is less effective compared to steady state update for replacement because it straightly changes the whole population chromosomes. Best parents from the previous generation are replaced by the new population. Hence, more iteration will be needed in order to find the optimum value. Steady state update is better because it eliminates the weak parents and conducts the competition between the previous generation parent and the new generation children. So, weak chromosomes can be replaced by the best chromosomes. Besides that, even better result can be produced and best parents from previous generation are still maintained in the population to be selected as parents in the next generation.

Based on the result in Table 3, it shows that the combination of techniques in GA1 model can produce good result and it is suitable to apply in the optimisation problems. Although the result for GA2 model is not very good, but the combination of methods for this model maybe is suitable to apply in other case studies.

5 Conclusion and Future Work

Based on the experiment result, Genetic Algorithm is proved to perform well in the optimisation problems but it highly depends on the techniques implemented. The techniques for each operation have shown different performance in obtaining the minimum and average values for benchmark functions. To conclude, different techniques applied in the operations will produce different results. Besides that, it will also affect the time taken for the model in producing result.

In order to improve the performance of GA, other operation techniques can be tested and maybe better result can be obtained. Besides that, new operation or enhancement can be applied on the technique and it might improve the result. Furthermore, hybrid with other soft computing methods can be implemented either inside or outside of GA operations. Probably this will improve the result obtained and more approximate towards the actual optimum value. In addition, same dataset can be used to be tested with different models and models that obtain the most minimum value can be obtained. This can also prove the performance of the models and techniques. The parameters setting can be adjusted so that most minimum value can be obtained for each technique and maybe termination criteria can be fulfilled faster.

Acknowledgement. This work is financed by UTM Zamalah. Authors would like to thank the anonymous referees for their valuable comments and suggestions.

References

1. Sivanandam, S.N., Deepa, S.N.: Introduction to Genetic Algorithm. Springer (2008)
2. Lim, S.P., Haron, H.: Surface Reconstruction Techniques: A Review. In: Artif. Intell. Rev., pp. 1–20. Springer (2012), doi:10.1007/s10462-012-9329-z
3. Vavak, F., Fogarty, T.C.: A Comparative Study of Steady State and Generational Genetic Algorithms for Use in Nonstationary Environments. In: Fogarty, T.C. (ed.) AISB-WS 1996. LNCS, vol. 1143, pp. 297–304. Springer, Heidelberg (1996)
4. Engelbrecht, A.P.: Computational Intelligence. John Wiley & Sons, Ltd. (2002)
5. Rajasekaran, S., Pai, G.A.V.: Neural Networks, Fuzzy Logic, and Genetic Algorithms Systhesis and Applications. Prentice-Hall of India Private Limited (2007)
6. Mansour, N., Awad, M., El-Fakih, K.: Incremental Genetic Algorithm. The International Arab Journal of Information Technology 3(1), 42–47 (2006)
7. Lin, C.T., Cheng, W.C., Liang, S.F.: Neural–Network–Based Adaptive Hybrid–Reflectance Model for 3–D Surface Reconstruction. IEEE Transactions On Neural Networks 16(6), 1601–1615 (2005)
8. Cheng, X., Wang, J., Wang, Q.: Leak–mending and Recruitment of Incomplete Points Data in 3D Reconstruction Based on Genetic Algorithm. In: Third International Conference on Natural Computation (ICNC 2007), pp. 259–263 (2007)
9. Saeedfar, A., Barkeshli, K.: Shape Reconstruction of Three–Dimensional Conducting Curved Plates Using Physical Optics, NURBS Modeling, and Genetic Algorithm. IEEE Transactions On Antennas And Propagation 54(9), 2497–2507 (2006)
10. Wang, S., Dhawan, A.P.: Shape–Based Reconstruction Of Skin Lesion For Multispectral Nevoscope Using Genetic Algorithm Optimization. In: 4th IEEE International Symposium on Biomedical Imaging: From Nano to Macro, ISBI 2007, pp. 488–491 (2007)
11. Sivaraj, R.: A Review Of Selection Methods In Genetic Algorithm. International Journal of Engineering Science and Technology (IJEST) 3(5), 3792–3797 (2011)
12. Goldberg, D.F., Deb, K.: A Comparative Analysis of Selection Schemes Used in Genetic Algorithms, Foundations of Genetic Algorithms, pp. 69–93. Morgan Kaufmann Publisher (1991)
13. Chudasama, C., Shah, S.M., Panchal, M.: Comparison of Parents Selection Methods of Genetic Algorithm for TSP. In: International Conference on Computer Communication and Networks CSI-COMNET-2011, Proceedings, pp. 85–87. International Journal of Computer Applications, IJCA (2011)
14. Othman, Z., Subari, K., Morad, N.: Job Shop Scheduling with Alternative Machines Using Genetic Algorithm. Jurnal Teknologi, 41(D) Dis. 2004, 67–78 (2004)
15. Molga, M., Smutnicki, C.: Test function for optimization needs, 1–43 (2005), http://www.zsd.ict.pwr.wroc.pl/files/docs/functions.pdf
16. Ortiz- Boyer, D., Hervás-Martínez, C., García-Pedrajas, N.: A Crossover Operator for Evolutionary Algorithms Based on Population Features. Journal of Artificial Intelligence Research 24, 1–48 (2005)
17. Hedar, A.R.: Test Functions for Unconstrained Global Optimization, http://www-optima.amp.i.kyoto-u.ac.jp/member/student/hedar/Hedar_files/TestGO_files/Page364.htm

A Runge-Kutta Method with Lower Function Evaluations for Solving Hybrid Fuzzy Differential Equations

Ali Ahmadian[1,⋆], Mohamed Suleiman[1], Fudziah Ismail[1],
Soheil Salahshour[2], and Ferial Ghaemi[3]

[1] Institute for Mathematical Research, Universiti Putra Malaysia,
43400 UPM, Serdang, Selangor, Malaysia
[2] Department of Mathematics, Mobarakeh Branch,
Islamic Azad University, Mobarakeh, Iran
[3] Institute of Advanced Technology, Universiti Putra Malaysia,
43400 UPM, Serdang, Selangor, Malaysia
ahmadian.hosseini@gmail.com,
{mohameds,fudizah}@science.upm.edu.my,
{soheilsalahshour,ferialghamei}@yahoo.com

Abstract. In this paper, we apply a Runge-Kutta method for solving first order fuzzy differential equations using lower number of function evaluations in comparison with classical Runge-Kutta method. It is assumed that the user will evaluate both f and f' readily instead of the evaluations of f only when solving hybrid fuzzy differential equation which enhance the order of accuracy of the solutions. Numerical example is provided which compares the new results with previous findings.

Keywords: Fuzzy ordinary differential equation, Hybrid system, Bede's Characterization Theorem, High order Runge-Kutta method, Seikkala derivative.

1 Introduction

The topics of fuzzy differential equations (FDEs) have been rapidly growing in recent years. In particular (FDEs) or (FIEs) appear in the study of synchronize hyperchaotic systems [26], control chaotic systems ([10], [12]) medicine ([1], [4]), bioinformatics and computational biology ([6],[7]). A thorough theoretical research of fuzzy first-order initial value problem was given by Kaleva [14] , Seikkala [25], Kloeden [15] and Wu [27]. Some applications of numerical and analytical methods such as the fuzzy Euler method [16], Adams-Bashforth, Adams-Moulton, Predictor-Corrector [2], Runge-Kutta [17], Laplace transform [24] in various kinds of FDE was presented to extend the implementation of Numerical and analytical methods for FODEs.

[⋆] Corresponding author.

A. Selamat et al. (Eds.): ACIIDS 2013, Part I, LNAI 7802, pp. 265–274, 2013.

Particularly, in recent years the use of hybrid fuzzy differential equations (HFDEs) has been increased more because it is a natural way to model control systems with embedded uncertainty that are capable of controlling complex systems which have discrete event dynamics as well as continuous time dynamics. (Pederson and Sambandham [18],[19],[20]) have presented the numerical solution of HFDEs by using the Euler and Runge-Kutta methods,respectively. Also (Prakash and Kalaiselvi [21]; Kima and Sakthivel [13]) have studied the predictor-corrector method for hybrid fuzzy differential equations and in [3] authors investigated the numerical solution of HFDEs by using the Euler method using characterization theorem and generalized differentiability.

The contribution of this paper is to use the extension of Bede's Characterization theorem for HFEDs which was presented by Pederson and Sambandham [18] to generalize derivatives. Then this results are used to solve HFDEs numerically by the new fourth order Runge-Kutta method.

A form of fourth Runge-Kutta with higher order derivative approximations will be explained in section 3 after preliminaries section . We study hybrid fuzzy differential equations using the concept of Characterization theorem in section 4. Numerical experiments are provided in Section 5 and compared with other methods. This is followed by a complete error analysis. At the end of the paper we present some conclusions.

2 Preliminaries

We give some definitions and introduce the necessary notation in this section which will be used throughout the paper. See for example ([28]).

We consider $\mathbb{R}$, the set of all real numbers. A fuzzy number is mapping $u : \mathbb{R} \to [0,1]$ with the following properties:

(a) u is upper semi-continuous,
(b) u is fuzzy convex, i.e., $u(\lambda x + (1-\lambda)y \geq min\{u(x), u(y)\}$ for all $x, y \in \mathbb{R}, \lambda \in [0,1]$,
(c) u is normal, i.e., $\exists x_0 \in \mathbb{R}$ for which $u(x_0) = 1$,
(d) $supp\ u = \{x \in \mathbb{R} | u(x) > 0\}$ is the support of the u, and its closure cl(supp u) is compact.

Let $\mathbb{E}$ be the set of all fuzzy numbers on r. The r-level set of a fuzzy number $u \in \mathbb{E}, 0 \leq r \leq 1$, denoted by $[u]_r$, is defined as

$$[u]_r = \begin{cases} \{x \in \mathbb{R} | u(x) \geq r\} & if\ 0 < r \leq 1 \\ cl(supp\ u) & if\ r = 0 \end{cases}$$

It is clear that r-level set of a fuzzy number is a closed and bounded interval $[\underline{u}(r), \overline{u}(r)]$, where $\underline{u}(r)$ denotes the left-hand endpoint of $[u]_r$ and $\overline{u}(r)$ denotes the right-hand endpoint of $[u]_r$. Since each $y \in \mathbb{R}$ can be regarded as a fuzzy number $\widetilde{y}$ is defined by

$$\tilde{y}(t) = \begin{cases} 1 & if \quad t = y \\ 0 & if \quad t \neq y \end{cases}$$

For $u, v \in \mathbb{E}$ and $\lambda \in \mathbb{R}$, the sum $u+v$ and the product $\lambda \odot u$ are defined by $[u+v]^\alpha = [u]^\alpha + [v]^\alpha$
$[\lambda \odot u]^\alpha = \lambda[u]^\alpha, \forall \alpha \in [0,1]$, where $[u]^\alpha + [v]^\alpha$ means that usual addition of two intervals (subsets) of $\mathbb{R}$ and $\lambda[u]^\alpha$ means the usual product between a scalar and a subset of $\mathbb{R}$.

The Hausdorff distance fuzzy numbers given by $D : \mathbb{E} \times \mathbb{E} \longrightarrow \mathbb{R}_+ \bigcup 0$,

$$D(u,v) = \sup_{r \in [0,1]} \max\{|\underline{u}(r) - \underline{v}(r)|, |\bar{u}(r) - \bar{v}(r)|\},$$

It is easy to see that D is a metric in $\mathbb{E}$ and has the following properties ([8])

(i) $D(u \oplus w, v \oplus w) = D(u,v), \quad \forall u, v, w \in \mathbb{E}$,
(ii) $D(k \odot u, k \odot v) = |k| D(u,v), \quad \forall k \in \mathbb{R}, u, v \in \mathbb{E}$,
(iii) $D(u \oplus v, w \oplus e) \leq D(u,w) + D(v,e), \quad \forall u, v, w \in \mathbb{E}$,
(iv) $(D, \mathbb{E})$ is a complete metric space.

Definition 1. *Let $f : \mathbb{R} \to \mathbb{E}$ be a fuzzy valued function. If for arbitrary fixed $t_0 \in \mathbb{R}$ and $\epsilon > 0, \delta > 0$ such that*

$$|t - t_0| < \delta \Rightarrow D(f(t), f(t_0)) < \epsilon,$$

f is said to be continuous.

Initially the H-derivative (Hukuhara differentiability) for fuzzy mappings was introduced by Puri and Ralescu [22] which is based on the H-difference sets, as follows.

Definition 2. *Let $x, y \in \mathbb{E}$.If there exists $z \in \mathbb{E}$ such that $x = y \oplus z$, then z is called the H-difference of x and y, and it is denoted by $x \ominus y$.*

In this paper, the sign " $\ominus$ " stands for H-difference, and also note that $x \ominus y \neq x + (-1)y$.

Definition 3. *Let $f : \mathbb{R} \to \mathbb{E}$ be a fuzzy function. We say f is differentiable at $t_0 \in \mathbb{R}$, if there exists an element $f^{'}(t_0) \in \mathbb{E}$ such that limits*

$$\lim_{h \to 0^+} \frac{f(t_0 + h) \ominus f(t_0)}{h} \quad and \quad \lim_{h \to 0^+} \frac{f(t_0) \ominus f(t_0 - h)}{h}$$

exist and are equal to $f^{'}(t_0)$. Here the limits are taken in the metric space $(\mathbb{E}, D)$, since we have defined $h^{-1} \odot (f(t_0) \ominus f(t_0 - h))$ and $h^{-1} \odot (f(t_0 + h) \ominus f(t_0))$.

Next we present Bede's Characterization theorem (let $\|.\|$ denote the usual Euclidean norm).

Theorem 1. *(Characterization Theorem) Let us consider the fuzzy initial value problem (FIVP) [5]*

$$\begin{cases} x' = f(t,x), \\ x(t_0) = x_0, \end{cases} \tag{1}$$

where $f : [t_0, t_0 + a] \times \mathbb{E} \to \mathbb{E}$ is such that *(i)* $[f(t,x)]^r = [\underline{f}^r(t,\underline{x},\overline{x}), \overline{f}^r(t,\underline{x},\overline{x})]$,

(ii) $\underline{f}^r$ *and* $\overline{f}^r$ *are equicontinuous (that is, for any* $\epsilon > 0$ *there is a* $\delta > 0$ *such that* $|\underline{f}^r(t,x,y) - \underline{f}^r(t,x,y)| < \epsilon$ *and* $|\overline{f}^r(t,x,y) - \overline{f}^r(t,x,y)| < \epsilon$ *for all* $r \in [0,1]$, *whenever* $(t,x,y), (t_1,x_1,y_1) \in [t_0, t_0 + a] \times \mathbb{R}^2$ *and* $\|(t,x,y) - (t_1,x_1,y_1)\| < \delta$ *and uniformly bounded on any bounded set,*

(iii) there exists an $L > 0$ *such that*
$|\underline{f}^r(t,x_1,y_1) - \underline{f}^r(t,x_2,y_2)| \leq L \max\{|x_2 - x_1|, |y_2 - y_1|\}$ *for all* $r \in [0,1]$,
$|\overline{f}^r(t,x_1,y_1) - \overline{f}^r(t,x_2,y_2)| \leq L \max\{|x_2 - x_1|, |y_2 - y_1|\}$ *for all* $r \in [0,1]$.

Then the FIVP (1) and system of ODEs

$$\begin{cases} (\underline{x}^r(t))' = \underline{f}^r(t,\underline{x}^r,\overline{x}^r) \\ (\overline{x}^r(t))' = \overline{f}^r(t,\underline{x}^r,\overline{x}^r) \\ \underline{x}^r(t_0) = (\underline{x_0^r}) \\ \overline{x}(t_0) = (\overline{x_0^r}) \end{cases} \tag{2}$$

are equivalent.

3 A Fourth-Order Runge-Kutta Method with Three Functional Evaluations

Consider the initial value problem

$$\begin{aligned} &y' = f(x,y), \\ &y(x_0) = y_0 \quad \textit{with } (x_0,y_0) \in D \end{aligned} \tag{3}$$

at which we assume that $f(x,y)$ has derivatives to the fourth order in domain D in $\mathbb{R}^{n+1}$ where $x \in \mathbb{R}, y \in \mathbb{R}^n$ and $(x,y) \in D$. Also we consider that

$$||f(x,y_1) - f(x,y_2)||_2 \leq L||y_1 - y_2||_2, \tag{4}$$

thus the problem (3) has a unique local solution.

Goeken and Johnson [11] introduced new terms involving higher order derivatives of f in the Runge-Kutta k_i terms $(i > 1)$ to achieve a higher order of accuracy without a corresponding increase in evaluations of f , but with the addition of evaluations or approximations of f' for third, fourth and fifth order method. The advantage of this method is that it has lower functional evaluation which improved the efficiency of method in comparison with classical Runge-Kutta and it can be applied for both autonomous and non-autonomous systems.

Treating the problem in autonomous form, For the fourth-order formula has the following form:

$$y_{n+1} = y_n + \frac{1}{6}k_1 + \frac{2}{3}k_2 + \frac{1}{6}k_3. \tag{5}$$

where

$$\begin{aligned}&k_1 = hf(y_n),\\ &k_2 = hf(y_n + a_{21}k_1 + ha_{22}f_y(y_n)k_1),\\ &k_3 = hf(y_n + a_{31}k_1 + a_{32}k_2 + ha_{33}f_y(y_n)k_1 + ha_{34}f_y(y_n)k_2).\end{aligned} \tag{6}$$

Specific nonzero constants, in the Runge-Kutta method with derivative approximations Goeken and Johnson (2000) for autonomous systems, are

$$b_1 = \frac{1}{6},\ b_2 = \frac{2}{3},\ b_3 = \frac{1}{6},\ a_{21} = \frac{1}{2},\ a_{22} = \frac{1}{8},\ a_{31} = -1,\ a_{32} = 2,\ a_{33} = -\frac{1}{2}.$$

3.1 Non-autonomous Derivations

If we proceed as above for $y' = f(x, y)$ we need to augment the terms involving $f_y k_i$ with hf_x . We can vary the terms used to match parameters. For example, if the equation is scalar, it is possible to use an f_y term in the x displacement rather than a f_x term in the y displacement. Since $f' = f_y f + f_x$ and since $f_y f$ is needed in the y displacement anyway, the exact computation of f_y is easier than the computation of f' Goeken and Johnson [11]. The following method uses f and f_y :

$$y_{n+1} = y_n + b_1k_1 + b_2k_2 + b_3k_3, \tag{7}$$

where

$$\begin{aligned}&k_1 = hf(x_n, y_n),\\ &k_2 = hf(x_n + hc_{21} + h^2c_{22}f_y, y_n + a_{21}k_1 + ha_{22}f_yk_1),\\ &k_3 = hf(x_n + hc_{31} + h^2c_{32}f_y, y_n + a_{31}k_1 + a_{32}k_2 + ha_{33}f_y(y_n)k_1 + ha_{34}f_y(y_n)k_2),\end{aligned} \tag{8}$$

where f_y is evaluated at (x_n, y_n).

The nonzero constant coefficients, in the Runge-Kutta method Goeken and Johnson [11] for non-autonomous systems, are

$$\begin{aligned}&b_1 = \tfrac{1}{6},\ b_2 = \tfrac{2}{3},\ b_3 = \tfrac{1}{6},\ c_{21} = a_{21} = \tfrac{1}{2},\ c_{22} = a_{22} = \tfrac{3}{32},\\ &a_{31} = -\tfrac{1}{2},\ a_{32} = \tfrac{3}{2},\ a_{33} = -\tfrac{11}{32},\ a_{34} = \tfrac{7}{32},\ c_{31} = 1,\ c_{32} = -\tfrac{1}{8}.\end{aligned} \tag{9}$$

4 Hybrid Fuzzy Differential Equation

In this paper, we consider the hybrid fuzzy differential equation

$$\begin{aligned}&x'(t) = f(t, x(t), \lambda_k(x_k)), \quad t \in [t_k, t_{k+1}], \quad k = 0, 1, 2, ...\\ &x(t_0) = x_0,\end{aligned} \tag{10}$$

where $\{t_k\}_{k=0}^{\infty}$ is strictly increasing and unbounded,x_k denotes $x(t_k)$, $f : [t_0, \infty) \times \mathbb{E} \times \mathbb{E} \to \mathbb{E}$ is continuous and each $\lambda_k : \mathbb{E} \to \mathbb{E}$ is continuous. A solution to (10) will be a function $x : [t_0, \infty) \to \mathbb{E}$ satisfying (10). For $k = 0, 1, 2, ...$, let $f_k : [t_k, t_{k+1}] \times \mathbb{E} \to \mathbb{E}$, where $f_k(t, x_k(t)) = f(t, x_k(t), \lambda_k(x_k))$.

The hybrid fuzzy differential equation (10) can be written as

$$x'(t)=\begin{cases} x_0'(t)=f(t,x_0(t),\lambda_0(x_0))=f_0(t,x_0(t)), & t_0\le t\le t_1,\\ x_1'(t)=f(t,x_1(t),\lambda_1(x_1))=f_1(t,x_1(t)), & t_1\le t\le t_2,\\ \vdots \\ x_k'(t)=f(t,x_k(t),\lambda_k(x_k))=f_k(t,x_k(t)), & t_k\le t\le t_{k+1},\\ \vdots \end{cases} \tag{11}$$

and a solution of (10) can be expressed as

$$x(t)=\begin{cases} x_0(t), & t_0\le t\le t_1,\\ x_1(t), & t_1\le t\le t_2,\\ \vdots \\ x_k(t), & t_k\le t\le t_{k+1},\\ \vdots \end{cases} \tag{12}$$

By using Bedes characterization theorem proposed by Bede [5], Pederson and Sambandham [18] generalized the following characterization theorem for hybrid fuzzy differential equation IVPs:

Theorem 2. *Consider the hybrid fuzzy differential equation IVP (10) expanded as (11) where for* $k=0,1,2,...,$ *each* $f_k:[t_k,t_{k+1}]\times\mathbb{E}\to\mathbb{E}$, *is such that*

(i) $[f_k(t,x)]^r=[(\underline{f}_k)^r(t,\underline{x}^r,\overline{x}^r),(\overline{f}_k)^r(t,\underline{x}^r,\overline{x}^r)]$,

(ii) $(\underline{f}_k)^r$ *and* $(\overline{f}_k)^r$ *are equicontinuous and uniformly bounded on any bounded set,*

(iii) there exists an $L_k>0$ *such that*
$|\underline{f}_k^r(t,x_1,y_1)-\underline{f}_k^r(t,x_2,y_2)|\le L_k\max\{|x_2-x_1|,|y_2-y_1|\}$ *for all* $r\in[0,1]$,
$|\overline{f}_k^r(t,x_1,y_1)-\overline{f}_k^r(t,x_2,y_2)|\le L_k\max\{|x_2-x_1|,|y_2-y_1|\}$ *for all* $r\in[0,1]$.

Then (10) and the hybrid system of ODEs

$$\begin{cases} ((\underline{x}_k)^r(t))'=\underline{f}_k^r(t,(\underline{x}_k)^r(t),(\overline{x}_k)^r(t))\\ ((\overline{x}_k)^r(t))'=\overline{f}_k^r(t,(\underline{x}_k)^r(t),(\overline{x}_k)^r(t))\\ (\underline{x}_k)^r(t_k)=(\underline{x}_{k-1})^r(t_k), \quad \textit{if } k>0,(\underline{x}_0)^r(t_0)=(\underline{x}_0)^r,\\ (\overline{x}_k)^r(t_k)=(\overline{x}_{k-1})^r(t_k), \quad \textit{if } k>0,(\overline{x}_0)^r(t_0)=(\overline{x}_0)^r, \end{cases} \tag{13}$$

are equivalent.

Proof. See [18]. □

5 Examples

Example 1. Consider the following hybrid fuzzy IVP,

$$\begin{cases} x'(t)=x(t)+m(t)\lambda_k(x(t_k)), \quad t\in[t_k,t_{k+1}],\ t_k=k, \quad k=0,1,2,...,\\ [x(0)]^r=[0.75+0.25r,\ 1.125-0.125r], \quad 0\le r\le 1, \end{cases} \tag{14}$$

where

$$m(t) = |\sin(\pi t)|, \qquad k = 0, 1, 2, ...,$$

$$\lambda_k(\mu) = \begin{cases} \hat{0} & if\, k = 0, \\ \mu & if\ k \in \{1, 2, ...\}. \end{cases}$$

From [18], for $t \in [1, 2]$ the exact solution of (26) is given by

$$\begin{aligned} \underline{x}(t; r) &= \underline{x}(1; r)\tfrac{\pi\cos(\pi t)+\sin(\pi t)}{\pi^2+1} + \tfrac{e^t}{e}x(1; r)\left(1 + \tfrac{\pi}{\pi^2+1}\right), \\ \overline{x}(t; r) &= \overline{x}(1; r)\tfrac{\pi\cos(\pi t)+\sin(\pi t)}{\pi^2+1} + \tfrac{e^t}{e}x(1; r)\left(1 + \tfrac{\pi}{\pi^2+1}\right). \end{aligned}$$

The results of the errors of the proposed RK4 formula, RK4 and Improved Euler method with $h = 0.5$ at $t = 1.5$ and $t = 2$are shown in Fig. 1 and Fig. 2. The exact and approximate solutions by Euler, the new RK4 and RK4 methods are compared and plotted at $t = 2$ in Fig. 3 . It is deduced that the results of the new Runge-Kutta method is very close to the exact solutions which confirm the validity and feasibility of this method.

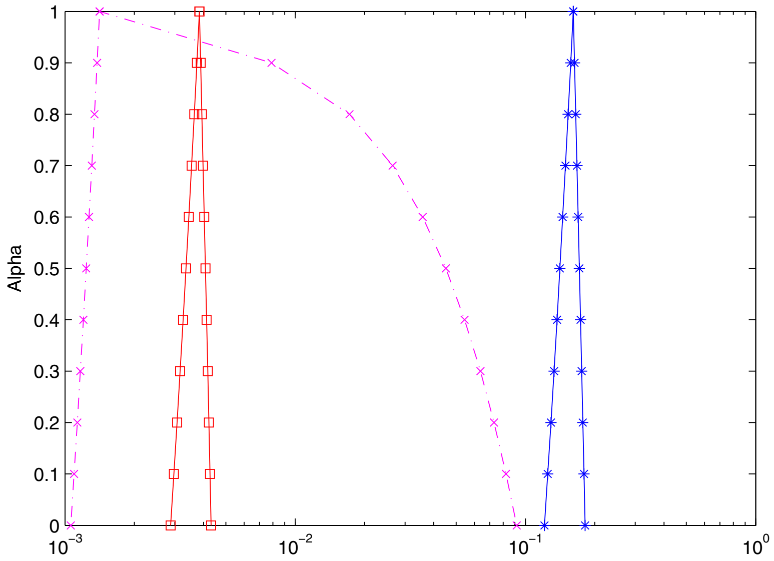


Fig. 1. The Absolute Error of: Presented Runge-Kutta method:□, Runge-Kutta method: x, Improved Euler method:*, Example 1 at $t = 1.5$

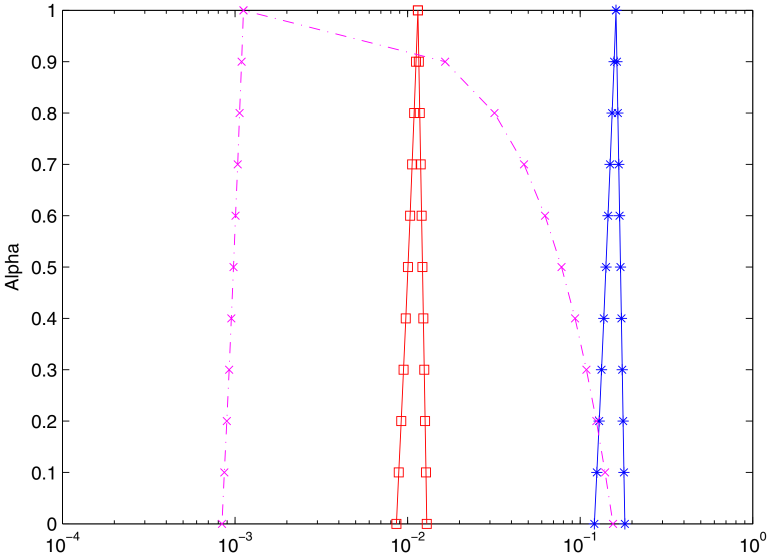


Fig. 2. The Absolute Error of : Presented Runge-Kutta method:□, Runge-Kutta method: x, Improved Euler method:*, Example 1 at $t = 2$.

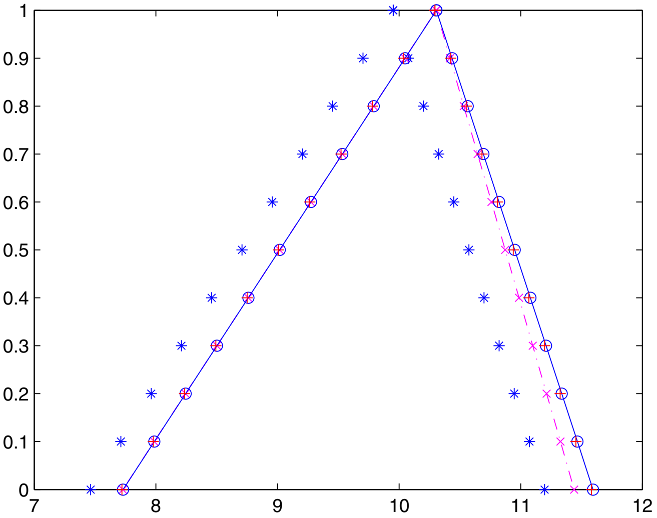


Fig. 3. Exact solution: O, Proposed Runge-Kutta method: +, Runge-Kutta method:-.x, Improved Euler method:*, Example 1 at $t = 2$.

6 Conclusion

In this paper a fourth order Runge-Kutta method has been presented which it has lower function evaluations in comparison with classical Runge-Kutta. A clear advantage of this method is that only three evaluation of function f and f' are required which for classical Runge-Kutta needs at least four evaluation functions. using the lower evaluation functions reduces the computational cost and is more attractive because the use of the value of f' sometimes is cheaper than evaluating f Results are comparable to Runge-Kutta solution of equal order, thus demonstrating our claim.

On the other side, the generalized characterization theorem was applied to transform the hybrid fuzzy ordinary differential equation to hybrid system of ODEs which means that any numerical method can be implemented for solving any kind of hybrid FODEs.

References

1. Abbod, M.F., Von Keyserlingk, D.G., Linkens, D.A., Mahfouf, M.: Survey of utilisation of fuzzy technology in medicine and healthcare. Fuzzy Set Syst. 120, 331–349 (2001)
2. Allahviranloo, T., Ahmady, N., Ahmady, E.: Numerical solution of fuzzy differential equations by predictor-corrector method. Information Sciences 177, 1633–1647 (2007)
3. Allahviranloo, T., Salahshour, S.: Euler method for solving hybrid fuzzy differential equation. Soft Comput. J. 15, 1247–1253 (2011)
4. Barro, S., Marn, R.: Fuzzy logic in medicine. Physica-Verlag, Heidelberg (2002)
5. Bede, B.: Note on "Numerical solutions of fuzzy differential equations by predictor corrector method". Information Sciences 178, 1917–1922 (2008)
6. Casasnovas, J., Rossell, F.: Averaging fuzzy biopolymers. Fuzzy Set Syst. 152, 139–158 (2005)
7. Chang, B.C., Halgamuge, S.K.: Protein motif extraction with neuro-fuzzy optimization. Bioinformatics 18, 1084–1090 (2002)
8. Dubios, D., Prade, H.: Towards fuzzy differential calculus-part3. Fuzzy Sets and Systems 8, 225–234 (1982)
9. Enright, W.H.: Second derivative multi-step methods for stiff ordinary differential equations. SIAM J. Numer. Anal. 11, 321–331 (1974)
10. Feng, G., Chen, G.: Adaptative control of discrete-time chaotic systems: a fuzzy control approach. Chaos, Solitons & Fractals 23, 459–467 (2005)
11. Goeken, D., Johnson, O.: Runge-Kutta with higher derivative approximations. Appl. Numer. Math. 39, 249–257 (2000)
12. Jiang, W., Guo-Dong, Q., Bin, D.: H_∞ variable universe adaptative fuzzy control for chaotic systems. Chaos, Solitons Fractals 24, 1075–1086 (2005)
13. Kima, H., Sakthivel, R.: Numerical solution of hybrid fuzzy differential equations using improved predictorcorrector method. Commun. Nonlinear Sci. Numer. Simulat. 17, 3788–3794 (2012)
14. Kaleva, O.: Fuzzy differential equations. Fuzzy Sets and Systems 24, 301–317 (1987)
15. Kloeden, P.: Remarks on Peano-like theorems for fuzzy differential equations. Fuzzy Set Syst. 44, 161–164 (1991)

16. Ma, M., Friedman, M., Kandel, A.: Numerical solution of fuzzy differential equations. Fuzzy Sets Syst. 105, 133–138 (1999)
17. Palligkinis, S.C., Papageorgiou, G., Famelis, I.T.: Runge-Kutta methods for fuzzy differential equations. Appl. Math. Comput. 209, 97–105 (2009)
18. Pederson, S., Sambandham, M.: Numerical solution of hybrid fuzzy differential equation IVPs by a characterization theorem. Information Sciences 179, 319–328 (2009)
19. Pederson, S., Sambandham, M.: The Runge-Kutta method for hybrid fuzzy differential equations. Nonlinear Anal. Hybrid Syst. 2, 626–634 (2008)
20. Pederson, S., Sambandham, M.: Numerical solution to hybrid fuzzy systems. Mathematical and Computer Modelling 45, 1133–1144 (2007)
21. Prakash, P., Kalaiselvi, V.: Numerical solution of hybrid fuzzy differential equations by predictor-corrector method. Int. J. Comput. Math. 86, 121–134 (2009)
22. Puri, M.L., Ralescu, D.: Differential for fuzzy function. J. Math. Anal. Appl. 91, 552–558 (1983)
23. Rosenbrock, H.H.: Some general implicit processes for the numerical solution of differential equations. Comp. J. 5, 329–330 (1963)
24. Salahshour, S., Allahviranloo, T., Abbasbandy, S.: Solving fuzzy fractional differential equations by fuzzy Laplace transforms. Commun. Nonlinear Sci. Numer. Simulat. 17, 1372–1381 (2012)
25. Seikkala, S.: On the fuzzy initial value problem. Fuzzy Sets and Systems 24, 319–330 (1987)
26. Zhang, H., Liao, X., Yu, J.: Fuzzy modeling and synchronization of hyperchaotic systems. Chaos, Solitons & Fractals 26, 835–843 (2005)
27. Wu, C., Song, S., Stanley Lee, E.: Approximate solution, existence and uniqueness of the Cauchy problem of fuzzy differential equations. J. Math. Anal. Appl. 202, 629–644 (1996)
28. Xu, J., Liao, Z., Hu, Z.: A class of linear differential dynamical systems with fuzzy initial condition. Fuzzy Sets Syst. 158, 2339–2358 (2007)

A Hybrid PSO-FSVM Model and Its Application to Imbalanced Classification of Mammograms

Hussein Samma[1], Chee Peng Lim[2,*], and Umi Kalthum Ngah[1]

[1] Imaging & Computational Intelligence Group (ICI),
School of Electrical and Electronic Engineering,
University of Science Malaysia, Malaysia
[2] Centre for Intelligent Systems Research, Deakin University, Australia
chee.lim@deakin.edu.au

Abstract. In this work, a hybrid model comprising Particle Swarm Optimization (PSO) and the Fuzzy Support Vector Machine (FSVM) for tackling imbalanced classification problems is proposed. A PSO algorithm, guided by the G-mean measure, is used to optimize the FSVM parameters in imbalanced classification problems. The hybrid PSO-FSVM model is evaluated using a mammogram mass classification problem. The experimental results are analyzed and compared with those from other methods. The outcomes positively demonstrate that the proposed PSO-FSVM model is able to achieve comparable, if not better, results for imbalanced data classification problems.

1 Introduction

Imbalanced classification problems occur when the instances in one class (i.e., the minority class) is very rare as compared with those in other classes (i.e., the majority classes). These problems exist in various fields, e.g. network intrusion detection [1] and breast cancer classification[2]. This study focuses on using a Fuzzy Support Vector Machine (FSVM) for undertaking imbalanced classification problems. While a number of computational intelligence models have been developed to undertake various problems in our previous studies [3-9], the models devised are not designed specifically for imbalanced learning and classification problems. As reported in the literature, Support Vector Machine (SVM)-based methods have been successfully applied to tackling imbalanced classification problems [10-20]. Generally these methods can be divided into two categories, i.e., integration of the SVM with data sampling techniques [11, 12, 14, 19] or modifications of the SVM [10, 13, 15-18, 20] to make it less sensitive to the class imbalanced problems. In the first category, sampling techniques (either over or under sampling) are first applied to the data level. Then, the SVM is used to process the sampled data.

Recently, Fuzzy Support Vector Machine (FSVM) [21] has been used for undertaking imbalanced learning problems [15]. Nonetheless, performances of an

[*] Corresponding author.

A. Selamat et al. (Eds.): ACIIDS 2013, Part I, LNAI 7802, pp. 275–284, 2013.

SVM-based model depends on the settings of its parameters, and inappropriate parameter selection leads to poor performances [22]. In this regards, Particle Swarm Optimization (PSO) has been used to optimize the SVM parameters [23, 24].

Motivated by the effectiveness of the FSVM in tackling imbalanced learning problems [15] and the usefulness of PSO in fine-tuning the SVM parameters [23, 24], this study proposes a hybrid PSO-FSVM model for undertaking imbalanced classification problems. The PSO algorithm is used to tune three FSVM parameters simultaneously, *viz.*, the penalty parameter (*c*) (which controls the trade-off between the model complexity and the training error rate); the kernel parameter (γ) of the kernel function (i.e., the Radial Basis Function); and the maximum fuzzy membership value (*r*). The applicability of the resulting PSO-FSVM model is demonstrated using a mammogram mass classification problem.

The organization of the paper is as follows. In Sections 2, and 3, the background pertaining to PSO and FSVM is given. Hybridization of PSO-FSVM is explained in Section 4. In Section 5, an experimental study to evaluate the effectiveness of PSO-FSVM using a mammogram mass classification problem is described. Concluding remarks and suggestions for further work are presented in Section 6.

2 Particle Swarm Optimization

PSO is a swarm-based optimization algorithm[25], which is inspired by the social behavior of organized colonies. A population in PSO is called a swarm, and it consists of particles which interact with each other to explore the search space. Each particle in the swarm is associated with two vectors, i.e., the velocity (*V*) and position (*X*) vectors, as follows:

$$X_i = [x_i^1, x_i^2, x_i^3, \dots\dots\dots\dots\dots, x_i^D] \tag{1}$$

$$V_i = [v_i^1, v_i^2, v_i^3, \dots\dots\dots\dots\dots, v_i^D] \tag{2}$$

where D represents the dimensions of the problem, *i* denotes the particle number in the swarm. Each particle is first initialized with random velocity and position vectors according to its solution range. During the search process, the velocity and the position vectors are updated, as follows:

$$V_i = w * V_i + c_1 * \text{rand}_1(\text{pBest} - X_i) + c_2 * \text{rand}_2(\text{gBest} - X_i) \tag{3}$$

$$X_i = X_i + V_i \tag{4}$$

where *w* is the inertia weight, c_1 and c_2 are the acceleration coefficients, $rand_1$ and $rand_2$ are two uniformly distributed random numbers in [0 1], *pBest* is the local best position achieved by the particle, and *gBest* is the global best position achieved by the swarm.

We adopt the procedure in [26] for setting the acceleration coefficients, i.e., the values of c_1 deceases from [2.5 to 0.5] and the values of c_2 increases from [0.5 to 2.5], respectively. As such, equations (5) and (6) are used:

$$c_1 = (0.5 - 2.5) * \frac{\text{current}_{\text{Iter}}}{\text{max}_{\text{Iter}}} + 2.5 \tag{5}$$

$$c_2 = 0.5 + \frac{\text{current}_{\text{Iter}}}{\text{max}_{\text{Iter}}} * (2.5 - .5) \tag{6}$$

where max_{Iter} is the maximum number of iterations, $\text{current}_{\text{Iter}}$ is the current iteration. The inertia weight parameter is a uniform random number $\epsilon[0\ 1]$, as suggested in [27].

3 Fuzzy Support Vector Machine

Originated from statistical learning theory, SVM aims to minimize structural risks by finding an optimal separating hyperplane which gives a low generalization error [21]. In SVM learning, minimization of the following function is considered

$$\min \frac{1}{2}\|w\|^2 + c \sum_{i=1}^{L} \xi_i \tag{7}$$

$$\text{subject to} \quad y_i(wx_i + b) \geq 1 \tag{8}$$

where ξ is the slack variable that measures the degree of misclassification, C is the penalty parameter that controls the trade-off between maximization of the margin and minimization of the classification error, b is the bias, and w is the weighted normal vector.

FSVM [21] is similar to SVM except that a fuzzy membership value, $m_{i.}$, is introduced, as follows

$$\min \frac{1}{2}\|w\|^2 + c \sum_{i=1}^{L} m_i \xi_i \tag{9}$$

$$\text{subject to} \quad y_i(wx_i + b) \geq 1 \tag{10}$$

The membership value, m_i, for each data point, x_i, in the training set is incorporated into the objective function (Eqn. 9). A smaller value of m_i makes the corresponding point x_i to be treated as less important, and vice versa. In this way, the cost of misclassification can be controlled by assigning lower membership values for less expensive classes and higher membership values for more expensive classes.

Another issue in FSVM is the determination of the membership value, $m_{i.}$ As stated in [21], the membership value can be computed as the distance between instance x_i and the class center. Let m_i+ to be the membership value for a more expensive class and m_i- is the membership value for a less expensive one, which are defined as follows:

$$m_i + = F(x_i) \tag{11}$$

$$m_i + = F(x_i) * \text{ratio} \tag{12}$$

where $ratio \in [\sigma \quad 1]$, and δ is a very small value selected by the user. As such, tuning the ratio value makes the FSVM classifier feasible in tackling imbalanced classification problems. .

4 A Hybrid PSO-FSVM Model

In this section, we proposed a hybrid PSO-FSVM model for undertaking imbalanced classification problems. The representation of the solution, fitness function, and procedure of the PSO-FSVM hybrid model are as follows.

(a) Solution representation

The FSVM with RBF kernels [15] $, k(x_i, x_j) = e^{-\gamma} \left\|x_i - x_j\right\|^2,$ is used in our proposed PSO-FSVM hybrid model. As shown in Table 1, each particle in the swarm consists of three real-valued variables, i.e., the penalty parameter *C*, the RBF kernel function parameter γ, and the ratio parameter that controls the maximum fuzzy membership value for the majority class instances, *r*.

Table 1. Parameters of a particle

penalty parameter (C)	RBF kernel parameter (γ)	Ratio parameter(r)
$x_{i,1}$	$x_{i,2}$	$x_{i,3}$

(b) Procedure

The main steps of the proposed PSO-FSVM model are illustrated in Fig 1, as follows.

1- Data pre-processing: The training data samples are scaled to lie within [-1 1].
2- Training: The FSVM classifier is trained using the training set.
3- Fitness evaluation: A validation set is used with the G-mean measure as the fitness function, and it is calculated for each particle in the swarm.
4- Stopping criteria: Two stopping conditions for the PSO algorithm are applied, i.e., either the maximum number of the PSO operations is met, or a G-mean value of 100 is achieved during the execution process.
5- PSO updating: The position and velocity vectors of each particle are updated according to equations (3) and (4).

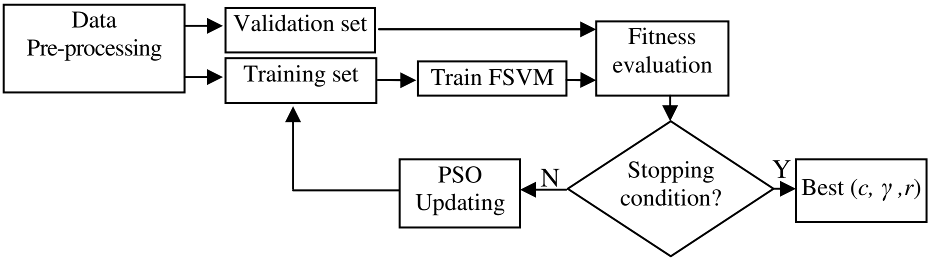


Fig. 1. PSO-FSVM model

(c) Fitness function

The G-mean measure is one of the assessment indicators that has been used to evaluate imbalanced data classification problems [15, 28, 29]. It is defined as follows.

$$\mathrm{G-mean} = \sqrt{\mathrm{Sensitivity} * \mathrm{Specificity}} \tag{13}$$

$$\mathrm{Sensitivity} = \frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FN}} \tag{14}$$

$$\mathrm{Specificity} = \frac{\mathrm{TN}}{\mathrm{TN}+\mathrm{FP}} \tag{15}$$

where TP is the true positive rate, FN is the false negative rate, TN is the true negative rate, and FP is the false positive rate.

5 An Experimental Study

The proposed PSO-FSVM model was evaluated using the MIAS mammogram data set [30]. A total of 92 abnormal Region of Interests (ROIs) (i.e., malignant and benign regions) and a total of 2090 normal ROIs were used in the experiment. As an example, a number of normal and abnormal ROI samples are shown in Fig 2. All ROI samples were re-scaled to 16x16 dimensions, and the intensity value was used to generate 256 feature vector lengths. The abnormal and normal ROIs were divided randomly into a training set and a testing set. For the training set, one third of the training samples were used for validation and the rest for training the FSVM classifier.

The PSO parameters were configured as in [26]. The size of the swarm and the maximum number of iterations were empirically chosen to be 20, and 1000, respectively. These values were found to be appropriate for this study by the trial-and-error method. The FSVM classifier with RBF kernels was used in the experiment. Both c and γ values were tuned within the logarithmic scale as recommended in [31], i.e., ln(c) =[-10,10], ln(γ) =[-10,10]. The last parameter, r, was tuned within the range r= [δ ,1],where δ is a very small value selected by the user (δ =$10^{\wedge(-6)}$).

(a) (b)

Fig. 2. (a) abnormal ROIs , (b) normal ROIs

5.1 Results and Discussions

In this study, the Center Composite Design (CCD) technique [32] was employed to analyze the effect of the PSO parameters (i.e., size of the swarm, S, and the number of iterations, N) on the FSVM accuracy rates. The CCD technique has been used to investigate the interaction between PSO parameters [33, 34]. Its basic idea is to set the value of each parameter at different levels (i.e. Low, Medium, and High), as shown in Table 2. Then, the combination of the experimented parameters at a different level is generated to study the interaction between the parameters, as shown in Fig 3. In this experiment, a total of five experiments were carried out using different levels for both S and N settings.

Table 2. Parameters and levels of the CCD experiment

Parameter	Level		
	Low (-1)	Medium (0)	High (+1)
S	10	20	40
N	100	1000	2000

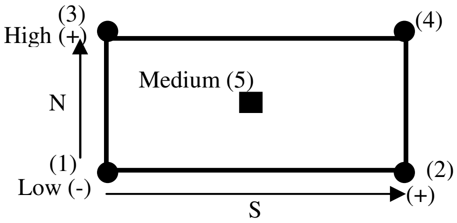


Fig. 3. A CCD experiment with two parameters and five points (i.e. one center and four corners)

Each experiment was repeated 10 times, and the average G-mean values were computed. The results are shown in Fig 4. As can be seen, the best G-mean value was produced by experiments 4 and 5. The parameter settings in experiment 5 (S=2, and N=1000) were used to compare PSO-FSVM with other methods in the next experiment.

A further experiment to compare the results of PSO-FSVM with those from PSO-SVM, FSVM, and SVM models, trained using the same data set, was conducted. In PSO-SVM, only two parameters were tuned by PSO, i.e. c and γ. In SVM and FSVM classifiers, three-fold cross validation was used to tune their parameters.

The experiment was repeated 10 times, and the mean and standard deviation of the sensitivity and specificity rates were computed. As shown in Table 3, the sensitivity rate and the G-mean value of PSO-FSVM are better than those from other SVM-based classifiers. However, its specificity rate is lower since the objective function used was maximization of the G-mean value, which was aimed to balance the true positive rate and the true negative rate.

Further analysis was conducted using the Receiver Operating Curve (ROC) as a graphical comparison method. As shown in Fig 5, PSO-FSVM produced a better ROC as compared with SVM-based classifiers. Moreover, the area under the ROC (AUC) was calculated, as in Table 3. The bootstrap method [35] was employed to assess the AUC values, with the significance level (α) set to 0.05 (i.e. 95% confidence level). As shown in Table 3, the *p-value* of the AUC measure is lower than α, which indicates that PSO-FSVM performed significantly better than the SVM-based FSVM classifiers (at the 95% confidence level).

Table 3. Comparison among different SVM classifiers

Model	Sensitivity (Std.dev)	Specificity (Std.dev)	G-mean (Std.dev)	AUC (Std.dev)	*p* value of bootstrap test
					AUC
MIAS dataset					
PSO-FSVM	**89.13** **(1.49e-014)**	90.21 (4.62e-002)	**89.67** **(2.29e-002)**	0.95 (1.31e-004)	-
PSO-SVM	49.35 (1.05)	**98.90** **(4.94e-002)**	69.86 (7.38e-001)	0.94 (4.59e-005)	0
SVM	51.09 (3.87)	98.68 (2.38e-001)	70.95 (2.63e+000)	0.94 (4.29e-003)	0
FSVM	51.30 (3.27)	98.74 (2.01e-001)	71.14 (2.21e+000)	0.94 (8.04e-003)	0.002

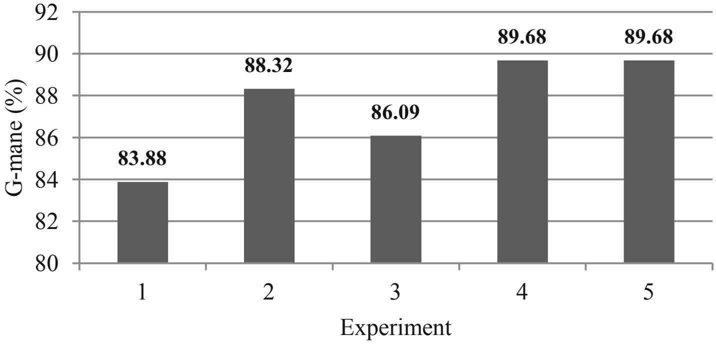


Fig. 4. Average G-mean values of PSO-FSVM

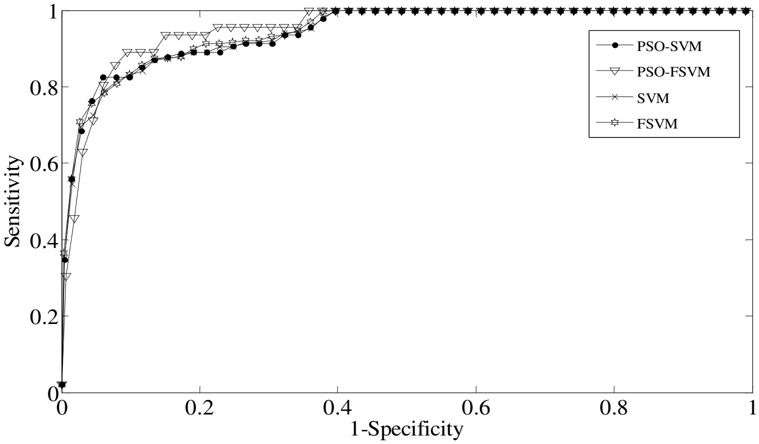


Fig. 5. Performance comparison with ROC with different PSO parameter settings

6 Summary

In this paper, a hybrid PSO-FSVM model has been proposed to undertake imbalanced classification problems. The effectiveness of PSO-FSVM has been evaluated using a mammogram mass classification problem. The results positively show that PSO-FSVM is able to produce good results, as compared to those from other methods.

While the performance of PSO-FSVM is encouraging, more experiments using imbalanced data sets are currently underway to further validate the usefulness of PSO-FSVM in undertaking imbalanced classification problems.

References

1. Vegard, E., Jonathan, V., Keith, P.: Enhancing network based intrusion detection for imbalanced data. Int. J. Know.-Based Intell. Eng. Syst. 12, 357–367 (2008)
2. Mazurowski, M.A., Habas, P.A., Zurada, J.M., Lo, J.Y., Baker, J.A., Tourassi, G.D.: Training neural network classifiers for medical decision making: The effects of imbalanced datasets on classification performance. Neural Networks 21, 427–436 (2008)
3. Chen, K., Lim, C., Lai, W.: Application of a Neural Fuzzy System with Rule Extraction to Fault Detection and Diagnosis. Journal of Intelligent Manufacturing 16, 679–691 (2005)
4. Eric, W.M.L., Lee, Y.Y., Lim, C.P., Tang, C.Y.: Application of a noisy data classification technique to determine the occurrence of flashover in compartment fires. Adv. Eng. Inform. 20, 213–222 (2006)
5. Goh, W.Y., Lim, C.P., Peh, K.K., Subari, K.: Application of a Recurrent Neural Network to Prediction of Drug Dissolution Profiles. Neural Computing & Applications 10, 311–317 (2002)
6. Chee-Peng, L., Jenn-Hwai, L., Mei-Ming, K.: A Hybrid Neural Network System for Pattern Classification Tasks with Missing Features. IEEE Trans. Pattern Anal. Mach. Intell. 27, 648–653 (2005)

7. Hua Tan, K., Peng Lim, C., Platts, K., Shen Koay, H.: An intelligent decision support system for manufacturing technology investments. International Journal of Production Economics 104, 179–190 (2006)
8. Chee Peng, L., Harrison, R.F.: Online pattern classification with multiple neural network systems: an experimental study. IEEE Transactions on Systems, Man, and Cybernetics, Part C: Applications and Reviews 33, 235–247 (2003)
9. Lim, C.P., Quek, S.S., Peh, K.K.: Prediction of drug release profiles using an intelligent learning system: an experimental study in transdermal iontophoresis. Journal of Pharmaceutical and Biomedical Analysis 31, 159–168 (2003)
10. Hsu, C.-C., Wang, K.-S., Chang, S.-H.: Bayesian decision theory for support vector machines: Imbalance measurement and feature optimization. Expert Systems with Applications 38, 4698–4704 (2011)
11. Liu, Y., An, A., Huang, X.: Boosting Prediction Accuracy on Imbalanced Datasets with SVM Ensembles. In: Ng, W.-K., Kitsuregawa, M., Li, J., Chang, K. (eds.) PAKDD 2006. LNCS (LNAI), vol. 3918, pp. 107–118. Springer, Heidelberg (2006)
12. Liu, Y., Yu, X., Huang, J.X., An, A.: Combining integrated sampling with SVM ensembles for learning from imbalanced datasets. Information Processing &amp; Management 47, 617–631 (2011)
13. Veropoulos, K., Campbell, C., Cristianini, N.: Controlling the Sensitivity of Support Vector Machines. In: the International Joint Conference on AI, pp. 55–60 (1999)
14. Kang, P., Cho, S.: EUS SVMs: Ensemble of Under-Sampled SVMs for Data Imbalance Problems. In: King, I., Wang, J., Chan, L.-W., Wang, D. (eds.) ICONIP 2006. LNCS, vol. 4232, pp. 837–846. Springer, Heidelberg (2006)
15. Batuwita, R., Palade, V.: FSVM-CIL: Fuzzy Support Vector Machines for Class Imbalance Learning. IEEE Transactions on Fuzzy Systems 18, 558–571 (2010)
16. Wu, G., Chang, E.Y.: KBA: kernel boundary alignment considering imbalanced data distribution. IEEE Transactions on Knowledge and Data Engineering 17, 786–795 (2005)
17. Zhao, Z., Zhong, P., Zhao, Y.: Learning SVM with weighted maximum margin criterion for classification of imbalanced data. Mathematical and Computer Modelling 54, 1093–1099 (2011)
18. Hwang, J.P., Park, S., Kim, E.: A new weighted approach to imbalanced data classification problem via support vector machine with quadratic cost function. Expert Systems with Applications 38, 8580–8585 (2011)
19. Yuchun, T., Yan-Qing, Z., Nitesh, V.C., Sven, K.: SVMs modeling for highly imbalanced classification. Trans. Sys. Man Cyber. Part B 39, 281–288 (2009)
20. Imam, T., Ting, K.M., Kamruzzaman, J.: z-SVM: An SVM for Improved Classification of Imbalanced Data. In: Sattar, A., Kang, B.-H. (eds.) AI 2006. LNCS (LNAI), vol. 4304, pp. 264–273. Springer, Heidelberg (2006)
21. Chun-Fu, L., Sheng-De, W.: Fuzzy support vector machines. IEEE Transactions on Neural Networks 13, 464–471 (2002)
22. Keerthi, S.S., Lin, C.-J.: Asymptotic behaviors of support vector machines with Gaussian kernel. Neural Computation 15, 1667–1689 (2003)
23. Guo, X.C., Yang, J.H., Wu, C.G., Wang, C.Y., Liang, Y.C.: A novel LS-SVMs hyper-parameter selection based on particle swarm optimization. Neurocomputing 71, 3211–3215 (2008)
24. Lin, S.-W., Ying, K.-C., Chen, S.-C., Lee, Z.-J.: Particle swarm optimization for parameter determination and feature selection of support vector machines. Expert Systems with Applications 35, 1817–1824 (2008)

25. Kennedy, J., Eberhart, R.: Particle swarm optimization. In: IEEE International Conference on Neural Networks, vol. 4, 1944, pp. 1942–1948 (1995)
26. Ratnaweera, A., Halgamuge, S.K., Watson, H.C.: Self-organizing hierarchical particle swarm optimizer with time-varying acceleration coefficients. IEEE Transactions on Evolutionary Computation 8, 240–255 (2004)
27. Sheng, C., Xia, H., Harris, C.J.: Particle Swarm Optimization Aided Orthogonal Forward Regression for Unified Data Modeling. IEEE Transactions on Evolutionary Computation 14, 477–499 (2010)
28. Gao, M., Hong, X., Chen, S., Harris, C.J.: A combined SMOTE and PSO based RBF classifier for two-class imbalanced problems. Neurocomputing 74, 3456–3466 (2011)
29. Xia, H., Sheng, C., Harris, C.J.: A Kernel-Based Two-Class Classifier for Imbalanced Data Sets. IEEE Transactions on Neural Networks 18, 28–41 (2007)
30. The Mini-MIAS Database of Mammograms, `http://peipa.essex.ac.uk`
31. Li, S., Tan, M.: Tuning SVM parameters by using a hybrid CLPSO-BFGS algorithm. Neurocomputing 73, 2089–2096 (2010)
32. Montgomery, D.C.: Design and analysis of experiments. Wiley (1997)
33. Pan, Q.-K., Fatih Tasgetiren, M., Liang, Y.-C.: A discrete particle swarm optimization algorithm for the no-wait flowshop scheduling problem. Computers &amp; Operations Research 35, 2807–2839 (2008)
34. Wang, C.-H., Lin, T.-W.: Improved particle swarm optimization to minimize periodic preventive maintenance cost for series-parallel systems. Expert Systems with Applications 38, 8963–8969
35. Efron, B.: Bootstrap Methods: Another Look at the Jackknife. The Annals of Statistics 7, 1–26 (1979)

QTCP: An Approach for Exploring Inter and Intra Protocols Fairness

Barkatullah Qureshi, Mohamed Othman[*], Shamala K. Subramaniam,
and Nor Asila Wati

Department of Computer Science and Information Technology, Universiti Putra Malaysia,
43400 UPM, Serdang, Selangor D.E., Malaysia
qureshi5939@gmail.com, mothman@fsktm.upm.edu.my

Abstract. TCP provides communication service however the network congestion is one of the core issues, it occurs when a link carry so much data, and effects on delay, packet loss rate and fairness. When the available bandwidth becomes very high, TCP, as tuned today, does not perform well. Several high-speed variants of TCP have been implemented to resolve the congestion issue , such as High-Speed TCP (HSTCP), Reno and New Reno TCP, Scalable TCP (STCP), Binary Increase Congestion Control TCP (BIC TCP), CUBIC TCP, Hamilton TCP (HTCP), Compound TCP and TCP Illinois. In this research a new Quick Transport Control Protocol (QTCP) has been designed to overcome the congestion issue. This is a new congestion control mechanism based on the modifications in the slow start phase of HSTCP and AIMD. QTCP performance evaluated by the experiment and compares the QTCP average throughput ratio, bandwidth link utilization and packet loss rate in inter and intra RTTs with inter protocols. We evaluated how the link can be fairly shared and achieve end-to-end throughput for each flow. Therefore a series of experiments conducted for the evaluation of multiple high-speed flows in different configurations. The results based on NS-2 simulator shows that QTCP provides better performance compared to other transport control protocols.

Keywords: Congestion Control, Fairness, Average Throughput, Bandwidth link utilization, Packet loss rate.

1 Introduction

The Sharing of data on the Internet and other similar networks is controlled by a suite of Internet protocols also known as Transmission Control Protocol/Internet Protocol (TCP/IP). TCP, being one of the protocols of the transport layer of the Internet Protocol Suite is also considered as one of its core protocols and handles most of the Internet data. Traditionally, while handling the data, TCP's congestion control performs well in most cases where the bandwidth is not extremely high, it shares the available bandwidth and offers fairness to millions of Internet users. However, when the available bandwidth becomes very high, TCP, as tuned today, does not perform well. This is because in

[*] The author is also an associate researcher at the Lab of Computational Science and Informatics, Institute of Mathematical Research (INSPEM), University Putra Malaysia.

A. Selamat et al. (Eds.): ACIIDS 2013, Part I, LNAI 7802, pp. 285–294, 2013.

the congestion avoidance phase, TCP takes a long time to increase the window size and cannot fully utilize the available bandwidth [1] and [2].Thus, the doors for the investigation of an optimal solution to this problem are still open and, consequently, in this article we propose our new approach, named the Quick Transport Control Protocol (QTCP) based on the modification of HSTCP. The QTCP takes a completely different approach in terms of congestion control over high-speed networks, and, hence, achieves improved throughput performance and fairness compared to existing high-speed TCP variants. The comparison of results based on the Network Simulator (NS-2) with several experimental configurations.

2 Related Work

To ameliorate congestion control issue, several high-speed variants of TCP have been implemented, such as High-Speed TCP (HSTCP) [3], Reno and New Reno TCP [4] and [5], Scalable TCP (STCP) [6], Binary Increase Congestion Control TCP (BIC TCP) [7], CUBIC TCP [8], Hamilton TCP (HTCP) [9], Compound TCP [10] and TCP Illinois [11]. FAST TCP is a delay-based approach and uses the RTTs for congestion measure; its throughput performance is significantly affected by the reverse traffic. Although the implementation and evaluation of all of these variants of TCP are available, none of these has yet been accepted as a single gold standard owing to the relative pros and cons of each variant [12].Therefore a new Quick Transport Control Protocol (QTCP) has been developed in this study.

3 Methodology

There is a plethora of algorithms aimed at the optimization of the standard TCP in terms of its congestion control mechanism to tune the requirements of applications that share large data over high-speed networks. Each of these algorithms has its own specific set of pros and cons and; to date, there is no single algorithm with the best optimal performance in all scenarios. This situation has led us to the development of yet another algorithm named the Quick Transport Control Protocol (QTCP). This algorithm implements a new congestion control mechanism based on the modifications in the slow start phase of HSTCP and AIMD. The modification is done in such a way that it provides significantly improved convenience, throughput, fairness and efficiency compared to most of the existing popular algorithms in the community. The parameters used in QTCP described in Table 1.

The QTCP growth function starts increasing the *cwnd* the same as other TCP schemes. The main function of QTCP is defined in equation (1), which has one parameter t, which represents the current time, or, we can say it represents the lifetime of a flow. QTCP depends on the parameter t that is being used in growth functions to compute the value used to increase the packet sending rate per RTT. In line one of the equation (1), it checks for the relation between *cwnd* and *ssthresh*; once the expression is satisfied it increases *cwnd* by 1, and the first packet will be sent to its destination. On successful acknowledgement of the sent packet, the process of increasing the sending rate per RTT and controlling the congestion window will be started. Hence, it will go to the second line, which increases

Table 1. Notation and Definition used in the QTCP Algorithm

Field	Description	Value
cwnd	Congestion window size.	-
ssthresh	Threshold value in TCP	-
t	Current/Elapsed Time	-
δ	Maximum increment in sending per RTT during α part.	3 seconds
Δ	Maximum increment in slow start	30 seconds
inc	A variable that holds the value, used to increase sending rate per RTT	-
$inc\alpha$	Increment in $inc\alpha$ in Slow Start.	4
$inc\beta$	Increment in Congestion Avoidance .	1
α	Life of a flow till second decrease in α part	
low_window	Low Window Parameter in Slow Start Phase.	-
updateInterval	Update Time Interval for $inc\alpha$	0.5 seconds
maxInc	Maximum Increase *cwnd* size in Congestion Avoidance Phase	10
$mode_{CA}$	A flag variable that switches QTCP to Congestion Avoidance mode of *cwnd* in Slow Start.	-
tq_lastupdate	Last Update Time.	-
decrease_factor	Window Decreasing Parameter when congestion happens	0.8

the value of *cwnd* linearly to increase the packet sending rate per RTT. The variable *inc* holds the value that is used to increase the packet sending rate per RTT. While the value of *inc* is computed by the $Qupdate(t)$ function defined in equation (2).

$$cwnd = Qtcp(t) = \begin{cases} cwnd + 1, & if\ cwnd < ssthresh \\ cwnd + inc = Qupdate(t), & otherwise \end{cases} \tag{1}$$

The function $Qupdate(t)$ defined in equation (2), which has one parameter inherited from the calling function $Qtcp(t)$, in line one of the function, checks the relation between *cwnd* and *low_window* (that is set to its default value 38). If the expression is satisfied, QTCP will exponentially increase the packet sending rate on each successful acknowledgement until the *cwnd* reaches to *low_window*. The exponential function $cwnd = 1/cwnd$.

$$cwnd = Qupdate(t) = \begin{cases} \frac{1}{cwnd}\ , & \text{if } cwnd < low\_window \\ Qincrease(t), & \text{otherwise} \end{cases} \tag{2}$$

Once *cwnd* hits the *low_window*, QTCP will start using the $Qincrease(t)$ function, which is defined in equation (3); again it has one parameter t, which is used to compute the value that increases the sending rate per RTT, as shown in Figure 1.

$$Qincrease(t) = \begin{cases} \alpha increase(t), & \text{if } mode_{CA} = FALSE \\ \beta increase(t), & \text{otherwise} \end{cases} \tag{3}$$

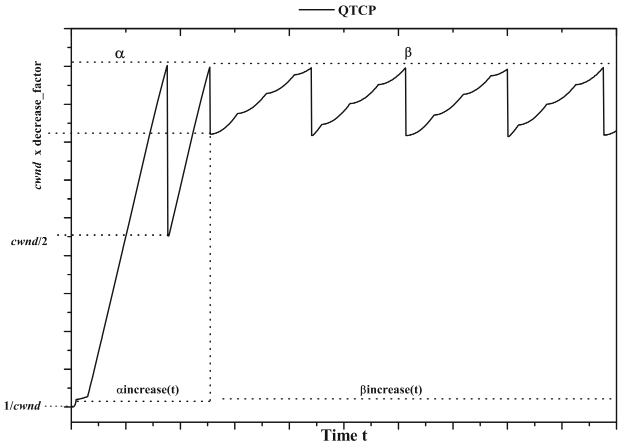


Fig. 1. Behaviour of the QTCP in slow-start and AIMD phase

QTCP divides the life of a flow into two parts, α or slow start: In this part *cwnd* must be increased faster with a certain speed so we can achieve the desired link utilization with minimum packet loss rate. β or congestion avoidance: In this part, the *cwnd* must be increased in such a way that promotes stability, utilization, fairness and minimum loss rate. Here in equation (3), $Qincrease(t)$ has two components; the first component $\alpha increase(t)$ is used to compute the value to increase the packet sending rate in α part. While $\beta increase(t)$ is used to compute the value to increase the packet sending rate in β or Congestion Avoidance part. On the α part we need to utilize the link capacity as much as we can; we need to achieve high link utilization with minimum packet loss rate. We designed a window growth function defined in equation (5), illustrated in Figure 2. It has one parameter t current time of a flow, in line one it checks the relation between t and Δ, which is set to its default value 3. If the expression is satisfied, it will return value of t as an increment in sending rate per RTT, until the value of t hits Δ. Once t hits the value of Δ, it will check the relation between *inc* and δ that is set to its default value 30. If the expression is satisfied, it will increase the value of *inc* by adding the value of $inc\alpha$ (that is set to its default value 4) to it, until *inc* hits the value of δ. Hence, although it will stop updating the value of *inc*, QTCP will continue to increase the sending rate by *inc*. The reason behind stopping the updating value of *inc* is to increase the stability and reduce the loss rate, as it will moderately increase the packet sending rate. In line three each time function saves current time t as a last update time $tq\_lastlupdate$ of *inc*.

QTCP keeps updating the sending rate until *cwnd* hits the maximum window size. Once it hits its maximum window size, the link will be congested and it will receive duplicate acknowledgements, which indicates that packets are being lost. For reducing congestion, QTCP decreases the window by half of *cwnd*, and a new threshold value will be set. This can be seen as follows in equation (4):

$$ssthresh = cwnd = cwnd/2 \tag{4}$$

Once the congestion is reduced, it stops receiving duplication acknowledgements; hence, QTCP will restart increasing the packet sending rate on each successful acknowledgement. The value of *inc* would not be updated, but its value is used to update the *cwnd* for increasing the packet sending rate, until it hits its maximum window size. Once it reaches the maximum window size, again the window will be decreased by the $decrease\_factor$, which is set to its default value 0.8, to reduce link congestion, and this time a flag variable $mode_{CA}$ will be set to true.

$$\alpha increase(t) = \begin{cases} t, & \text{if } t < \Delta \\ inc + inc\,\alpha, & \text{if } inc < \delta \\ tq\_lastupdate = t & \end{cases} \tag{5}$$

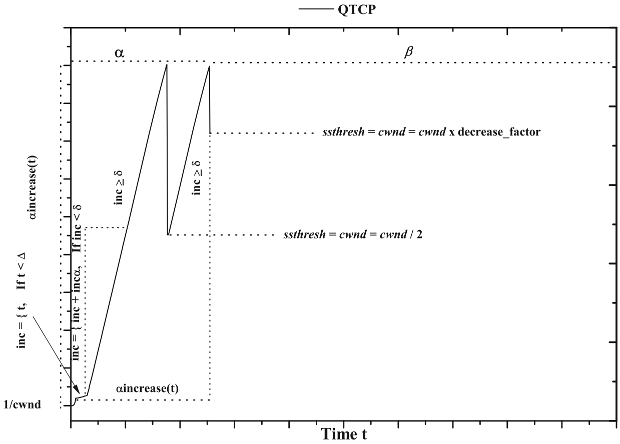


Fig. 2. The window growth function of QTCP α in slow-start phase

The flag variable $mode_{CA}$ switches QTCP to congestion avoidance phase. In congestion avoidance or β, the value for increasing packet sending rate per RTT, will be computed by β $increase(t)$, defined in equation (6), as shown in Figure 3.

$$\beta increase(t) = \begin{cases} inc + inc\beta,\ tq\_lastupdate = t, & \text{if } t - tq\_lastupdate \geq updateInterval \\ 1, & \text{if } inc > maxInc \end{cases} \tag{6}$$

In line one of $\beta increase(t)$ function, it computes the new value for increasing the sending rate, by adding *inc* and $inc\beta$ (which is set to its default value 1 after a certain interval called *updateInterval* which is set to its default value 0.5 seconds. After every 0.5 seconds $\beta increase(t)$ updates the value of *inc* by adding $inc\beta$ to it. This gives a new value to the increased sending rate. While QTCP keeps increasing the packet sending rate on

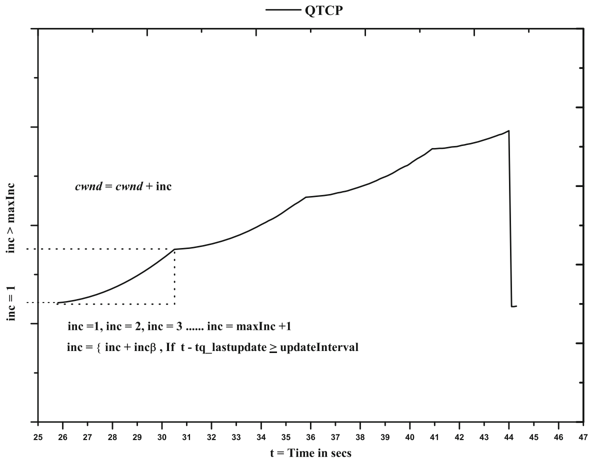


Fig. 3. The window growth function of QTCP α in slow-start phase

each successful acknowledgement by adding *inc* to *cwnd*, and waits for $updateInterval$ and again updates *inc* by adding $inc\beta$. In line two it checks if *inc* hits its maximum value *maxInc* (which is set to its default value 10) then it resets it to 1. This process will be continued until *cwnd* hits its maximum window size, when *cwnd* hits its maximum window size then the window will be decreased by the $decrease\_factor$ for reducing congestion to avoid packet loss. Once the window is decreased, again QTCP will keep increasing the sending rate by computing *inc* using $\beta increase(t)$ in the absence of congestion, QTCP continues this process of avoiding congestion, maintaining Stability, maintaining fairness and achieving utilization until the flow completes its data transfer.

4 Performance Evaluation

QTCP evaluated on the basis of the average throughput ratio, packet loss rate and link utilization between inter protocols with different RTTs in the given experiment. Simple topologies, like a dumbbell with one congested link, are sufficient to study many traffic properties.

4.1 Experimental Setup

In this section, the performance of the QTCP with different RTTs is analysed. The main attempt of the configuration is to compare the QTCP average throughput ratio, bandwidth link utilization and packet loss rate in inter and intra RTTs with inter protocols. The topology used in these experiments is shown in Figure 4.

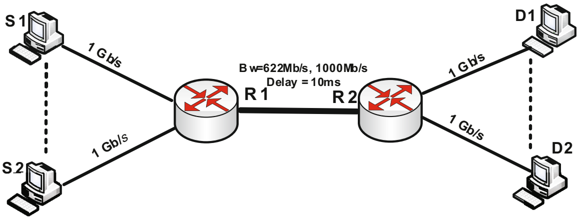


Fig. 4. Network topology used for connections with different RTTs

Dumbbell topology with two routers R1-R2 is located at the bottleneck between the two end points. The maximum bandwidth of the bottleneck routers is set in 622Mb/s and 1Gb/s, and the delay is set at 10ms. The dotted line shows the connection link between the sender/receiver nodes with the capacity of 1Gb/s. The bottleneck buffer size of the senders (S1-S2) and receivers (D1-D2) sides are fixed at 1555 packets. The start time of different flows is set between 0.1 and 0.5 seconds for each simulation, each time simulation run for 1000 seconds. In this section, inter protocols QTCP simulations are analysed with different RTTs. To determine the behaviour of different connections with different RTTs, the round trip time of connection one fixed at 160ms while the round trip time of connection to varied at 24ms, 40ms, 60ms, 80ms and 160ms. The average throughput ratios of each competing connection between 24ms RTT to 160ms RTT are shown in Figure 5(a) and Figure 5(b). The bandwidth utilization is restricted to 622Mb/s in Figure 5(a), which shows QTCP has a better average throughput ratio compared to all the inter protocols except HTCP. The average throughput ratio of HTCP slightly decreases whenever the RTTs increase. The QTCP bandwidth utilization with 1000Mb/s in Figure 5(b) reveals a better average throughput ratio with all inter protocols except HTCP and FAST. The FAST TCP has a slightly smaller average throughput ratio than HTCP.

Figure 6(a) and Figure 6(b) shows the results of QTCP inter protocol packet loss rate with different RTTs, and 622Mb/s and 1000Mb/s bandwidths, respectively. In Figure 6(b) it reveals that the FAST TCP packet loss rate is slightly higher at 80ms RTT. STCP has high packet loss rate, as shown in Figure 6(a) and Figure 6(b). In this evaluation, the QTCP shows the lower packet loss rate with all other inter protocols except FAST TCP and STCP, therefore it is found that QTCP is more compatible with all other high-speed TCP protocols.Figure 7(a) and Figure 7(b) shows the results of QTCP inter protocol link utilization with different RTTs, and 622Mb/s and 1000Mb/s bandwidths, respectively. Figure 7(a) confirms that QTCP has best link utilization, which shows the better share of resources with all inter protocols. As compared to the large RTTs, the short RTTs provide better link utilization results by up to 99.88%, and 76.53% in large RTTs. Figure 7(b) indicates that QTCP better link utilization with all the inter protocols except HTCP and FAST TCP. The HTCP link utilization gradually decreased, when the RTTs increased. The FAST TCP has slightly smaller link utilization at 60ms RTT.

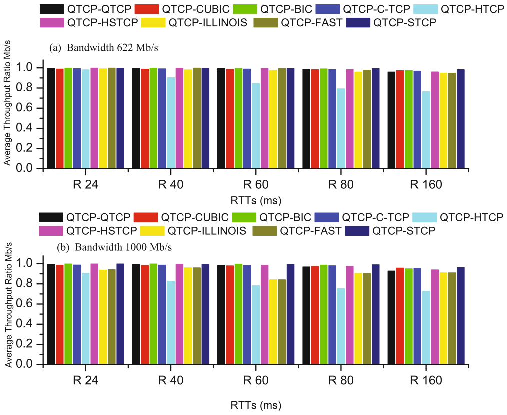


Fig. 5. Inter protocols throughput ratio with different RTTs

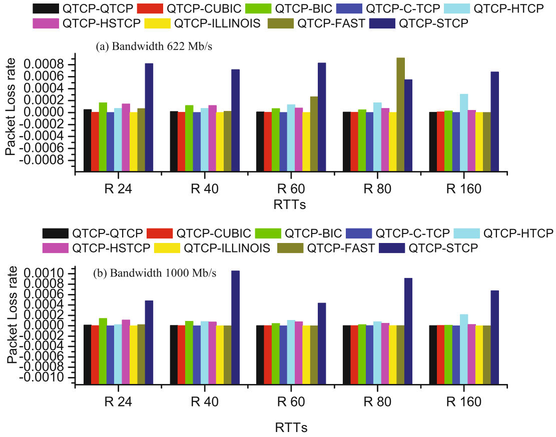


Fig. 6. Inter protocols packet loss rate with different RTTs

4.2 Fairness

We computed the fair share link metric by using equation (7), which describes the Jain fairness index [13]. A series of experiments for the evaluation of multiple high-speed flows in different configurations is shown in Table 2.

$$F = \frac{(\sum_{i=1}^{n} \bar{x}_i)^2}{n \sum_{i=1}^{n} \bar{x}_i{}^2} \qquad (7)$$

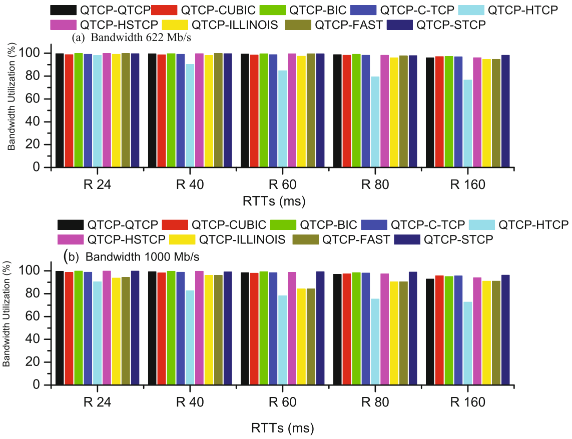


Fig. 7. Inter protocols bandwidth utilization with different RTTs

Table 2. QTCP fairness evaluated with other high-speed protocols

	Bandwidth Mb/s				
Protocols	200	400	600	800	1000
QTCP-QTCP	1.00	1.00	1.00	0.99	0.98
QTCP-CUBIC	0.99	0.99	0.98	0.99	0.99
QTCP-BIC	1.00	1.00	1.00	0.99	0.99
QTCP- C-TCP	0.98	0.99	0.99	0.99	0.98
QTCP-HTCP	0.99	0.99	0.99	0.99	0.99
QTCP-HS	0.99	0.99	1.00	0.99	0.98
QTCP-ILLINOIS	1.00	1.00	1.00	0.99	0.98
QTCP-FAST	1.00	1.00	1.00	0.99	0.98
QTCP-STCP	0.90	0.90	0.95	0.96	0.95

5 Conclusion

The QTCP, a new variant of high-speed TCP, was proposed and studied in this article. The most significant finding that emerged from this study is that the QTCP achieves improved fairness in terms of the multiple flows sharing the link and it also offers high throughput in terms of complex network resources compared to most of the existing high-speed TCP variants. The inter and intra protocol fairness evaluation over different RTTs, verified that the QTCP can effectively utilize the link capacity with low packet losses and excellent throughput ratio. Consequently, it is recommended that the future trials should analyse the parallel flows that use multiple effective TCP connections. It is assumed that as the parallel TCP would aggregate, the throughput would increase, the loss rate would decrease and, hence, even better throughput would be achieved.

Acknowledgements. The research was supported by the Malaysian Ministry of Higher Education, Fundamental Research Grant (FRGS) 01-11-09-734FR.

References

1. Xue-zeng, P., Fan-jun, S., Yong, L., Ling-di, P.: CW-HSTCP: fair TCP in high-speed networks. Journal of Zhejiang University Science 7(2), 172–178 (2006)
2. Fan-jun, S., Xue-Zeng, P., Jie-Bing, W., Zhengi, W.: An algorithm for reducing loss rate of high-speed TCP. Journal of Zhejiang University Science 7 (supp.II), 245–251 (2006)
3. Floyd, S., Ratnasamy, S., Shenker, S.: Modifying TCP's congestion control for high speeds, Technical note (May 2002), `http://www.icir.org/floyd/papers/hstcp.pdf`
4. Zhang, Y., Lemin, L., Wang, S.: Improving Reno and New-Reno's performances over OBS networks without SACK. International Journal of Electronics and Communications 63(4), 294–303 (2009)
5. Floyd, S., Henderson, T., Gurtov, A.: The NewReno modification to TCP's fast recovery algorithm (April 2004), `http://www.faqs.org/rfcs/rfc3782.html`
6. Kelly, T.: Scalable TCP: Improving performance in high-Speed wide area networks. ACM SIGCOMM Computer Communication Review 33(2), 83–91 (2003)
7. Xu, L., Harfoush, K., Rhee, I.: Binary increase congestion control for fast, long distance networks. In: Proceeding of NETWORKING 2006, pp. 476–487 (2006)
8. Rhee, I., Xu, L.: CUBIC: A new TCP-friendly high-speed TCP variant. In: Proceeding of PFLDnet (2005)
9. Shorten, R., Leith, D.: H-TCP: TCP for high-speed and long-distance networks. In: Proceeding of Second International Workshop on Protocols for Fast-long Distance Networks (2004)
10. Tan, K., Song, J., Zhang, Q., Sridharan, M.: A compound TCP approach for high-speed and long distance networks. In: Proceeding of 25th IEEE INFOCOM International Conference on Computer Communications (2006)
11. Liu, S., Basar, T., Srikant, R.: TCP-Illinois: A loss- and delay-based congestion control algorithms for highspeed networks. In: Proceeding of 25th IEEE INFOCOM International Conference on Computer Communications, Performance Evaluation, vol. 65, pp. 417–440 (2008)
12. Weigle, M.C., Sharma, P., Freeman, J.: Performance of competing high-speed TCP flows. In: Proceedings of the IFIP Networking, Coimbra, Portugal (2006)
13. Chiu, D., Jain, R.: Analysis of the increase/decrease algorithm for congestion avoidance in computer networks. Computer Networks and ISDN 17(1) (June 1989)

Analyzing Hemagglutinin Genes of Human H5N1 Virus by Linear Neighborhood Embedding

Wei-Chen Cheng

Institute of Statistical Science, Academia Sinica
Taiwan, Republic of China
r93108@csie.ntu.edu.tw

Abstract. This work proposes using linear neighborhood embedding to visualize the distribution of viral genes in a two-dimensional space. The gradient descent algorithm for the embedding is described and the behavior of the algorithm is analyzed.

Keywords: LRE, DNA, influenza, population, immunological distance.

1 Introduction

Gene sequences have been studied by metric methods. Cheng [2,5] projected the DNA sequences onto a space, spanned by three orthogonal axes [12]. The projection uses the matrix calculated from the length of longest common sequences (LCS) among pairwise genes, of which the inverse indicates the dissimilarity of the two sequences. The computational complexity for calculating the distance is proportional to the product of the two sequence lengths. The length of the LCS has a reciprocal relation to the edit distance, which manipulates one gene sequence to match the other gene by three equally weighted operations: substitution, insertion and deletion. The edit distance had also been used in Isomap for constructing three-dimensional projections of Human genetic sequences that maintain the global geometry [2,6]. The Isomap [10] employs shortest path to approximate the geodesic distance among data, and projects sequences onto the low-dimensional space that best preserve the covariance of the geodesic distance matrix. The length of the shortest path between two sequences is symmetric. The length is calculated from the adjacent matrix with edit distance among sequences. In the latest work, Cheng [2] used a distance-invariant self-organizing map and the inverse length of LCS to construct a set of low-dimensional representations that the zooming pairwise distances among data in high-dimensional space can be preserved. The subtypes of influenza A virus, such as H5N1 and H1N1, were analyzed [7].

Paccanaro and Hinton proposed the linear relational embedding [8], a method for learning distributed representations of items from the data consisting of instances of relations among given items. The binary relation is either on or off.

A. Selamat et al. (Eds.): ACIIDS 2013, Part I, LNAI 7802, pp. 295–303, 2013.

In the situation that the distance measurement is unreliable and only the relationships among items need to be considered, the relational embedding method is suitable to handle this kind of problems. The relations in the triplet clause are represented by linear transformations and items are described by vectors. The transformation is directional, $R(a,b) \neq R(b,a)$, and is different from the metric methods, which adopt the distance matrix for expressing and analyzing data relations. The distance measurement of the metric methods is symmetric, $d(a,b) = d(b,a)$. Therefore, the linear relational embedding algorithm is flexible to generate the vectors to represent items in Euclidean space without the explicit restriction of distance. The method had been applied to solve the family tree problem and number problem. In this work, we describe a linear neighborhood embedding method for studying gene sequences.

2 Method

Suppose we have a set of sequences $V = \{\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_N\}$. The number of sequences in the set $|V|$ is N. We also have the set of relations of the rank $R = \{R_1, R_2, \ldots, R_K\}$. The R_1 denotes the first place, the R_2 denotes the second, and so on. The K is the number of relations. Considering the K nearest neighbors of all sequences, we can convert any distance matrix into a set of triplets, $(\mathbf{v}_1, R_1, \mathbf{v}_{I_{11}})$, $(\mathbf{v}_1, R_2, \mathbf{v}_{I_{12}})$, $(\mathbf{v}_1, R_3, \mathbf{v}_{I_{13}})$, ..., $(\mathbf{v}_1, R_K, \mathbf{v}_{I_{1K}})$, $(\mathbf{v}_2, R_1, \mathbf{v}_{I_{21}})$, ..., $(\mathbf{v}_2, R_K, \mathbf{v}_{I_{2K}})$, ... The I_{ij} is the index refer to the jth nearest sequence to the sequence $\mathbf{v}_i$. The overall number of the triplets is $N \times K$. The sensible likelihood function is as follows [8]:

$$\begin{aligned} G &= \sum_{i=1}^{N}\sum_{j=1}^{K} g_{ij} \\ &= \sum_{i=1}^{N}\sum_{j=1}^{K} \log \frac{e^{-\left(R_j\mathbf{v}_i - \mathbf{v}_{I_{ij}}\right)^T\left(R_j\mathbf{v}_i - \mathbf{v}_{I_{ij}}\right)}}{\sum\limits_{c=1}^{N} e^{-(R_j\mathbf{v}_i - \mathbf{v}_c)^T(R_j\mathbf{v}_i - \mathbf{v}_c)}} . \end{aligned} \tag{1}$$

Let t_{ij} denotes the numerator $e^{-\left(R_j\mathbf{v}_i - \mathbf{v}_{I_{ij}}\right)^T\left(R_j\mathbf{v}_i - \mathbf{v}_{I_{ij}}\right)}$ in (1) and s_{ij} denotes the denominator $\sum_{c=1}^{N} e^{-(R_j\mathbf{v}_i - \mathbf{v}_c)^T(R_j\mathbf{v}_i - \mathbf{v}_c)}$. One of the ways to adjust the matrices in R and the vectors in V is the gradient descent method. The gradient of G with respect to a sequence $\mathbf{v}_k$ is the summation of the derivatives of all components:

$$\sum_{i=1}^{N}\sum_{j=1}^{K} \frac{\partial g_{ij}}{\partial \mathbf{v}_k} = \sum_{i=1}^{N}\sum_{j=1}^{K} \frac{1}{t_{ij}}\frac{\partial t_{ij}}{\partial \mathbf{v}_k} - \frac{1}{s_{ij}}\frac{\partial s_{ij}}{\partial \mathbf{v}_k} . \tag{2}$$

To calculate the derivative in (2) with respect to the elements in V, we discuss four conditions of (1). The first condition is that three indices i, I_{ij}, and k refer to the same sequence $\mathbf{v}_i = \mathbf{v}_{I_{ij}} = \mathbf{v}_k$. The second condition is that they refer to different sequences $\mathbf{v}_i \neq \mathbf{v}_{I_{ij}} \neq \mathbf{v}_k$. The third and fourth conditions are that

the k equals to one of the other two different indices: $\mathbf{v}_i = \mathbf{v}_k$ or $\mathbf{v}_{I_{ij}} = \mathbf{v}_k$ but $\mathbf{v}_i \neq \mathbf{v}_{I_{ij}}$. The V is updated in the batch mode, which means the vector representations of all the sequences move simultaneously at each iteration.
Conditions:

1) $\mathbf{v}_i = \mathbf{v}_{I_{ij}} = \mathbf{v}_k$: The derivative in Equation (2) is

$$\begin{aligned}\frac{\partial g_{ij}}{\partial \mathbf{v}_k} &= -2\left[R_j^T R_j \mathbf{v}_k - \left(R_j + R_j^T\right)\mathbf{v}_k + \mathbf{v}_k\right] \\ &\quad + \frac{2}{s_{ij}}\left[R_j^T R_j \mathbf{v}_k - \left(R_j + R_j^T\right)\mathbf{v}_k + \mathbf{v}_k\right] e^{-(R_j\mathbf{v}_k - \mathbf{v}_k)^T(R_j\mathbf{v}_k - \mathbf{v}_k)} \\ &\quad + \frac{2}{s_{ij}}\sum_{c\neq k}\left[R_j^T R_j \mathbf{v}_k - R_j^T \mathbf{v}_c\right] e^{-(R_j\mathbf{v}_k - \mathbf{v}_c)^T(R_j\mathbf{v}_k - \mathbf{v}_c)} \end{aligned} \tag{3}$$

2) $\mathbf{v}_i \neq \mathbf{v}_{I_{ij}} \neq \mathbf{v}_k$: The change of $\mathbf{v}_k$ only affect the denominator of G.

$$\frac{\partial g_{ij}}{\partial \mathbf{v}_k} = \frac{2}{s_{ij}}\left(-R_j\mathbf{v}_i + \mathbf{v}_k\right) e^{-(R_j\mathbf{v}_i - \mathbf{v}_k)^T(R_j\mathbf{v}_i - \mathbf{v}_k)} \tag{4}$$

3) $\mathbf{v}_i \neq \mathbf{v}_{I_{ij}}$ and $\mathbf{v}_i = \mathbf{v}_k$: The derivative in Equation (2) is

$$\begin{aligned}\frac{\partial g_{ij}}{\partial \mathbf{v}_k} &= -2\left[R_j^T R_j \mathbf{v}_k + R_j^T \mathbf{v}_{I_{ij}}\right] \\ &\quad + \frac{2}{s_{ij}}\left[R_j^T R_j \mathbf{v}_k - \left(R_j + R_j^T\right)\mathbf{v}_k + \mathbf{v}_k\right] e^{-(R_j\mathbf{v}_k - \mathbf{v}_k)^T(R_j\mathbf{v}_k - \mathbf{v}_k)} && (5) \\ &\quad + \frac{2}{s_{ij}}\sum_{c\neq k}\left[R_j^T R_j \mathbf{v}_k - R_j^T \mathbf{v}_c\right] e^{-(R_j\mathbf{v}_k - \mathbf{v}_c)^T(R_j\mathbf{v}_k - \mathbf{v}_c)} && (6) \end{aligned}$$

4) $\mathbf{v}_i \neq \mathbf{v}_{I_{ij}}$ and $\mathbf{v}_{I_{ij}} = \mathbf{v}_k$: The derivative in Equation (2) is

$$\begin{aligned}\frac{\partial g_{ij}}{\partial \mathbf{v}_k} &= 2\left[R_j\mathbf{v}_i - \mathbf{v}_k\right] \\ &\quad + \frac{2}{s_{ij}}\left[-R_j\mathbf{v}_i + \mathbf{v}_k\right] e^{-(R_j\mathbf{v}_i - \mathbf{v}_k)^T(R_j\mathbf{v}_i - \mathbf{v}_k)} \end{aligned} \tag{7}$$

We have obtained $\frac{\partial G}{\partial \mathbf{v}_k}$ and then calculate the derivative of (1) with respect to R_k, $k \in \{1, \ldots, K\}$,

$$\begin{aligned}\frac{\partial g_{ij}}{\partial R_k} &= -2\left[R_k\mathbf{v}_i\mathbf{v}_i^T + \mathbf{v}_{I_{ij}}\mathbf{v}_i^T\right] \\ &\quad + \frac{2}{s_{ij}}\sum_{c}\left[R_k\mathbf{v}_i\mathbf{v}_i^T - \mathbf{v}_c\mathbf{v}_i^T\right] e^{-(R_k\mathbf{v}_i - \mathbf{v}_c)^T(R_k\mathbf{v}_i - \mathbf{v}_c)}. \end{aligned} \tag{8}$$

The value of V and R starts from small random numbers in the range $[-1, 1]$ in this work.

3 Simulations and Results

3.1 Frog Data with Immunological Distance

Traditional species groups are primarily defined by similarities in external morphology, such as the osteological features. The study of frog species by Case [1] includes using the immunological distances among nine frog (Rana) species. The phylogenetic tree is built by Fitch and Margoliash [3] method from the immunological distances. The immunological distance is related to the number of amino acid sequence differences in antigens. If immunological distance depended on only the difference in amino acid sequence between homologous and heterologous antigens, the distance between two proteins X and Y should be the same whether the antisera used were anti-X or anti-Y. Reciprocal tests provide a measure of how well the differences in amino acid sequences are estimated.

Table 1. The converted neighborhood relations

Species of Frog	Abbreviation	Ranked Neighbors
R. aurora	(a)	b,m,c,p,tem
R. boylii	(b)	c,m,p,a,tem
R. cascadae	(c)	b,m,p,a,tem
R. muscosa	(m)	b,c,a,p,tem
R. pretiosa	(p)	b,c,m,a,tem
R. tarahumarae	(t)	pip,cat,p,m,b
R. pipiens	(pip)	t,m,cat,p,c
R. catesbeiana	(cat)	m,c,p,pip,t
R. temporaria	(tem)	c,m,p,b,a

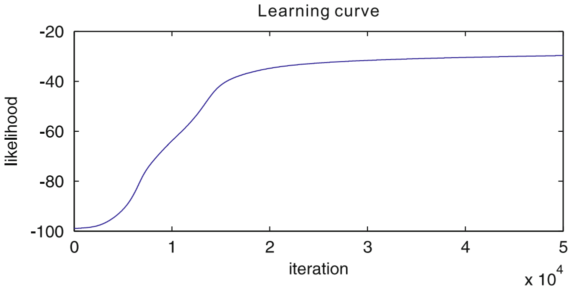


Fig. 1. The plot records the value of the likelihood defined in (1) during the training

Therefore, the immunological distance for analysis is calculated from reciprocal tests among the albumins of the nine species of ranid frogs. We convert the immunological distance matrix into 45 triplets (a, R_1, b), (a, R_2, m), (a, R_3, c), (a, R_4, p)...(tem, R_5, a) by setting the K to be 5. The detailed information is

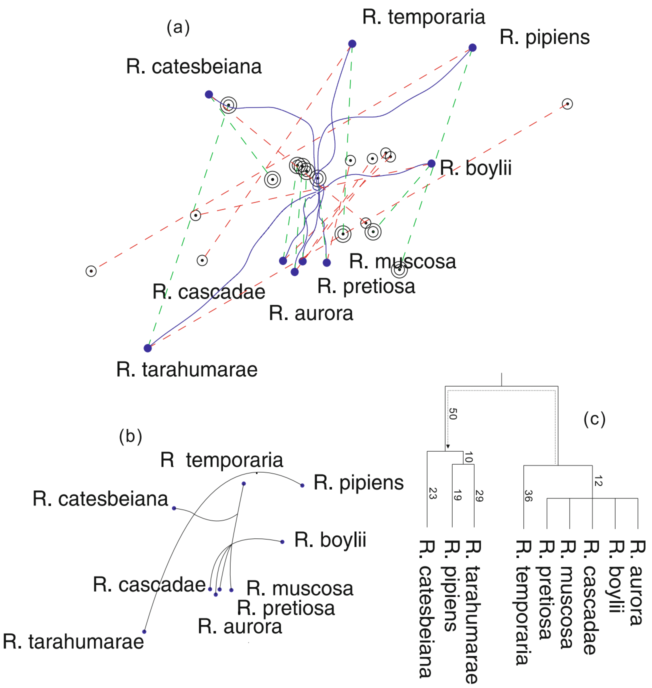


Fig. 2. The plot in the top shows the updating trajectory of the coordinates, representing the frogs, in the space. The red dashed line indicates the resulted location after applying the rank 1 relation R_1 on the frog. The green dashed line indicates the resulted location after applying R_2.

listed in Table 1. The first column lists the species name of the frogs, the second shows their abbreviations, and the third shows their neighbors. The values of (1) during the fifty thousand iterations are plotted in Figure 1. The training is converged after twenty thousand iterations. The vector representations and relational matrices embedded in a two-dimensional space are plotted in Figure 2. The blue curve in Figure 2(a) show the trajectory of the vector representations of the sequences during the updating process. The converged locations are ploted in Figure 2(b). The tree in Figure 2(c) is transcribed from the research in [1] for comparison. The linkages (branches) are attached to the vertices in Figure 2(b). We see that the right trunk of the tree matches the two-dimensional representations and the left trunk is dispersed. The vector fields of the relational matrices, R_1 and R_2, are visualized in Figure 3.

3.2 Influenza with Hamming Distance

Influenza belongs to the family Orthomyxoviridae and is a single negative-strand RNA virus. Its genome is not a single piece of nucleic acids but contains seven or eight pieces of segmented negative-sense RNA. Hemagglutinin (HA) and neuraminidase (NA) are the two glycoproteins on the outside of the viral particles.

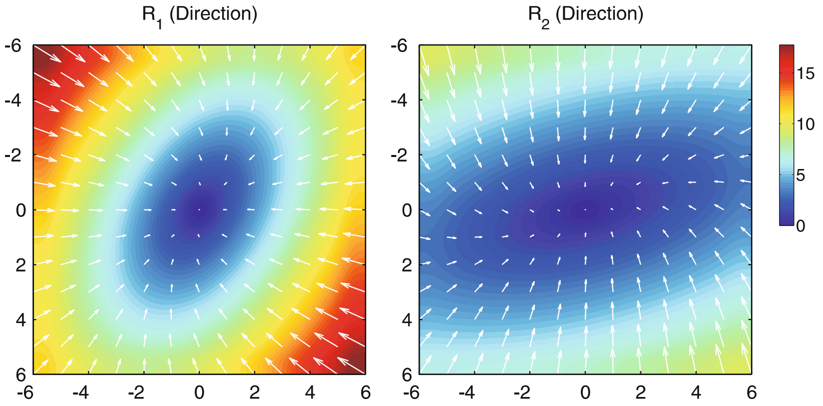


Fig. 3. The vector flow of the relations R. The background color shows the distance of displacement. The directional force of R_1 and R_2 plotted on the grid in the range $[-6, 6]$.

They are the main target of vaccine research. The 201 amino acid sequences of Hemagglutinin were downloaded from the website of NCBI [11] and aligned altogether.

The multiple-sequence alignment is based on the program Clustal [4]. Basically, the algorithm consists of three main stages. The first stage is to calculate a distance matrix by separately aligning all pairs of sequences. The second stage is to estimate a guide tree from the distance matrix and the third stage is to progressively align the sequences according the branching order in the tree. Each alignment of all pairs of sequences is by full dynamic programming whose computational complexity is proportional to the product of the lengths of the sequence pair. Three parameters have to be given. The weight matrix for the dynamic programming of protein sequences is given by Gonnet series. Two gap penalties, one is for opening a new gap and the other is for extending an existing gap, are given to be 10 and 0.1. After the distance matrix is calculated, the next step is to construct the guide tree by Neighbour-Joining method [9] according to the matrix. The last step builds a series of pairwise alignments to align enlarged size of sequence groups. There is 53.61% identity in the aligned sequences after performing the alignment to all of them. Considering the statement in [11] and due to the sequences are all from the same type of virus, the group of the sequences is not highly divergent (larger than 25 ∼ 30% identity). Therefore, the progressive approach is suitable for aligning influenza sequences. The alignment algorithm has dealt with that the gaps in protein sequence occur more often between the secondary structural elements of α-helices and β-strands than within. And it's rare to have a gap opening at a place less than eight residues away from existing gaps but often to have gaps in the hydrophilic regions. Hence the gap penalty varies with the location along the sequence. Having the aligned sequences, we used Hamming distance to calculate the difference among them and to build the distance matrix that has biological basis.

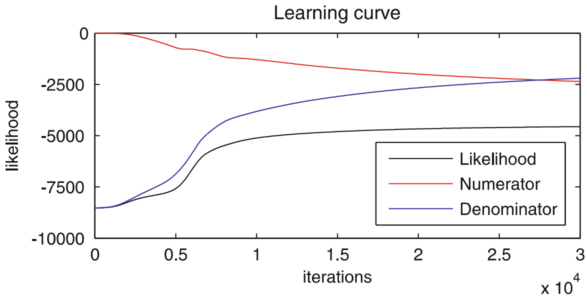


Fig. 4. The curves record the value of likelihood function during updating

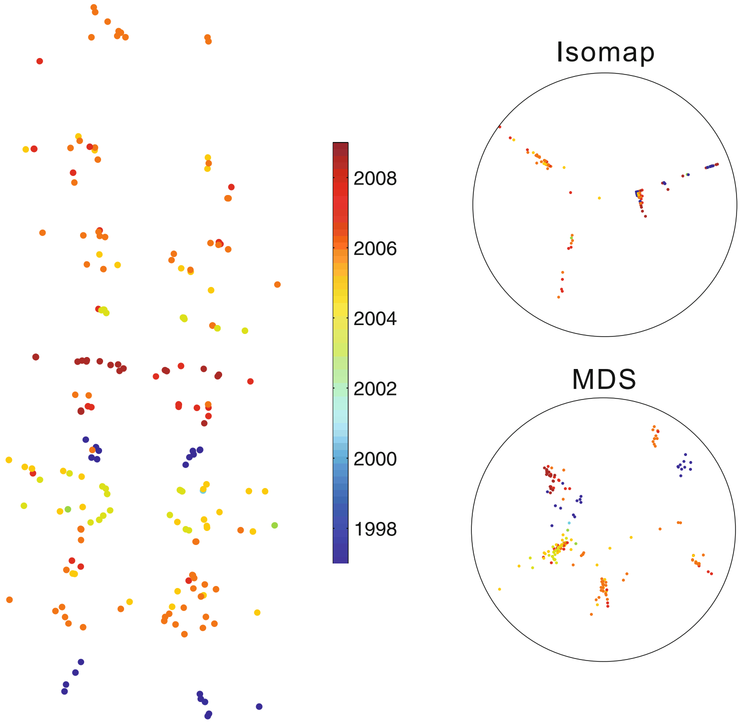


Fig. 5. The distribution of virus that found by linear neighborhood embedding is shown with two other methods

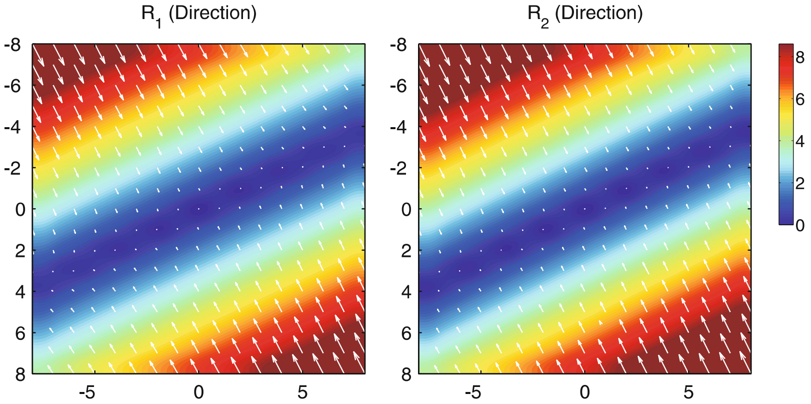


Fig. 6. The vector field of the relations found by linear neighborhood embedding. The background color indicates the real length of the displacements.

The length of the aligned sequences was 582 amino acids. There being 1608 triplets were extracted from the Hamming distances among all aligned sequences. The value of K was set to be 8. We tried three different initial values for the vectors V and relations R. The minimal converged value of the likelihood function at the 30000 iteration was -4560.7. The curve, which records the changes of the likelihood function (1), is shown in Figure 4. The numerator is calculated by $\sum_{i=1}^{N} \sum_{j=1}^{K} \log t_{ij}$ and the denominator is by $\sum_{i=1}^{N} \sum_{j=1}^{K} \log s_{ij}$ in (1). The distribution resulted from linear neighborhood embedding is shown in Figure 5. The color of the dots marks the year information. The result of Isomap [10] and MDS is also plotted for comparison. These two methods, one is nonlinear and the other is linear, use distance information to calculate the embedding.

4 Summary

In this paper, we introduced the development of the analysis of genetic sequence by manifold learning and proposed linear neighborhood embedding, which is derived from linear relational embedding, to analyze gene sequences. The results compare different methods and show the properties of linear neighborhood embedding. The embedding not only generates the representations for virus, but also the vector field for accomplishing the neighboring relations.

Acknowledgement. This work was supported by National Science Council under project NSC101-2811-M-001-082.

References

1. Case, S.M.: Biochemical Systematics of Members of the Genus Rana Native to Western North America. Syst. Zool. 27, 299–311 (1978)
2. Cheng, W.-C.: Visualizing Human Genes on Manifolds Embedded in Three-Dimensional Space. In: Pan, J.-S., Chen, S.-M., Nguyen, N.T. (eds.) ACIIDS 2012, Part II. LNCS, vol. 7197, pp. 421–430. Springer, Heidelberg (2012)
3. Fitch, W.M., Margoliash, E.: Construction of Phylogenetic Trees. Science 155, 279–284 (1967)
4. Higgins, D.G., Sharp, P.M.: CLUSTAL: a Package for Performing Multiple Sequence Alignment on a Microcomputer. Gene 73, 237–244 (1988)
5. Krejcar, O., Janckulik, D., Motalova, L.: Complex Biomedical System with Biotelemetric Monitoring of Life Functions. In: Proceedings of the IEEE Eurocon, pp. 138–141 (2009)
6. Krejcar, O., Janckulik, D., Motalova, L.: Complex Biomedical System with Mobile Clients. In: Dossel, O., Schlegel, W.C. (eds.) WCMPBE 2009, vol. 25/5, pp. 141–144. Springer, Munich (2009)
7. Liou, C.-Y., Cheng, W.-C.: Visualization of Influenza Protein Segment HA in Manifold Space. In: Nguyen, N.T., Le, M.T., Świątek, J. (eds.) ACIIDS 2010. LNCS, vol. 5990, pp. 150–158. Springer, Heidelberg (2010)
8. Paccanaro, A., Hinton, G.E.: Learning Distributed Representations by Mapping Concepts and Relations into a Linear Space. In: Langley, P. (ed.) ICML 2000, pp. 711–718. Stanford University, Morgan Kaufmann Publishers, San Francisco (2000)
9. Saitou, N., Nei, M.: The Neighbor-joining Method: A New Mothed for Reconstructing Phylogenetic Trees. Mol. Biol. Evol. 4, 406–425 (1987)
10. Tenenbaum, J.B., de Silva, V., Langford, J.C.: A Global Geometric Framework for Nonlinear Dimensionality Reduction. Science 290, 2319–2323 (2000)
11. Thompson, J.D., Higgins, D.G., Gibson, T.J.: CLUSTAL W: Improving the Sensitivity of Progressive Multiple Sequence Alignment through Sequence Weighting, Position-specific Gap Penalties and Weight Matrix Choice. Nucleic Acids Res. 22, 4673–4680 (1994)
12. Torgerson, W.S.: Multidimensional Scaling, I: Theory and Method. Psychometrika 17, 401–419 (1952)

The VHO Project: A Semantic Solution for Vietnamese History Search System

Dang-Hung Phan[1] and Tuan-Dung Cao[2]

[1] Institution of Information Technology,
Vietnam Academy of Science and Technology
danghung@ioit.ac.vn
[2] School of Information and Communication Technology
Hanoi University of Science and Technology
dungct@soict.hut.edu.vn

Abstract. This paper presents the process of building a semantic search application in the field of the Vietnamese history. In particular, the design of ontology in the field of history will come up with three main issues: the historical ontology construction with Vietnamese history books or documents as an input, managing change entities and the storage method of time entities in historical ontology. The experimental result of this article is the VHO web-based application which allows users to ask semantic questions about historical objects.

Keywords: Semantic web, historical ontology, information extraction.

1 Introduction

For a field that has characteristics of a vast of knowledge like history, requirements to build a smart lookup function for historical websites is essential. Currently, search functions in almost all well-known historical websites, like [16] in Vietnam and [17] in the world, are classical methods based on the keyword search method in historical articles. This method often provides low accuracy, especially when we use a number of complex queries, including detailed questions about historical objects, such as question 1.1**:** Find generals who againsted the Mongol enemy?.

We can overcome this disadvantage by storing detailed information about historical objects in a relational database system. However, this method doesn't cover the case when users ask questions which force the search system to find out the inference results, called interference question. Let's consider an example question 1.2: Find leaders defeated the Mongol enemy?. Clearly, question 1.2 doesn't enter in the same way with question 1.1 we set out, but by inference we can see that the question 1.2 is the specific case of questions 1.1, i.e. "general" is the specific concept of "leader" and "defeat the enemy" is the specific action of "against the enemy".

Semantic technology is an effective solution that we could build such smart search functions, which can cover main disadvantages of the keyword search and the

A. Selamat et al. (Eds.): ACIIDS 2013, Part I, LNAI 7802, pp. 304–313, 2013.

relational database search methods. The use of semantic technology platform is used to build up an information system named Vietnamese History Ontology, abbreviates VHO.

Related Work

The predicate-object search method in VHO semantic search has been learned from the semantic search system for tourism [4], but it still has its own specific characteristics because of the field of history. The specific characteristics of a semantic search application domain in the field of Vietnamese history are learned and improved through the process of reference to the historical ontology construction projects in Europe [5, 6, 7] and the ontology construction projects for storing historical events described in Franklin De Roosevelt library (United States)'s documents, called the Pearl Harbor [2] project. In particular, the project [2] has given us the challenges to build a historical ontology, there are three issues: the historical ontology construction with Vietnamese history books or documents as an input, managing changes in ontology historical entities and storage method for time entities, based on which we have developed methods of storing ontology to solve the problem mentioned above. Along with the design process, ontology VHO also be compared and additional concepts from historical ontologies of projects [5, 6, 7].

Outlines

The content of this paper will describe features of semantic search system on the field of Vietnamese history VHO. Specifically, after Introduction, section 2 will introduce ontology design methodology, describes the main problems to be solved in the process of build VHO ontology. Next, section 3- system architecture will introduce tools used to build this semantic application. Experimental results of this semantic search application will be given in Section 4. In this section, we will also evaluate the accuracy of VHO search results.

2 Building VHO Historical Ontology

The building process of VHO historical ontology was based on 7 steps to create ontology introduced in [8]. To build a semantic application that can meet needs of the lookup users in a specific field, we have to find and handle specific points of that field. This section will describe 5 steps that reflect specific points of the Vietnamese history field.

2.1 Scope of Historical Domain

Ontology VHO is designed within the input key words (concept) appear with great frequency in a set of three Vietnamese history documents: 2 historical books [14], [15] and a set of 63 historical articles collected from [16]. They are three historical documents that written and edited by historical experts. These materials are highly appreciated for the accuracy of information about historical objects. There are 4 main kind of objects in these document: historical characters, organizations, events and places. When building ontology, we will put child concepts of these four types. Therefore, we use the top-down design methodology to build ontology VHO. Specifically, we will build four branches concepts with the highest level of 4 historical subjects mentioned.

2.2 Concept Information Extraction

2.2.1 Term Extraction

In this project, JvnTextPro ([9]) helps us to classify the vocabulary of the three historical documents of Vietnam. Analytical result of term extraction is a set of vocabulary includes 18,247 nouns, and in which we have selected a list of about 130 nouns appear with great frequency, they are specific concepts of historical events, historic sites, historical organizations, and historical characters. The followings table is the list of greatest frequency concepts of concept “historical characters”.

Table 1. Term extrator for concept “Historical character”

Historical characters (Vietnamese)	**English meaning**	**Number count**
Vua	King	3031
Quan lai	Official who serves the King	1530
Tuong quan	General	693
Cong chua	Princess	138

2.2.2 Taxonomy Extraction

By the semantic dictionary WordNet ([10]), we will know position of input words in the wordnet tree vocabulary. However, in some case there are only vocabulary in Vietnamese history and can not get definition in WordNet as "Thuong thu", "Thi lang" (these words are a kind of “Quan lai”) ... In this case, we will assign this word as the child concept of nearest specific word that can be found by WordNet. For example, “Thuong thu” is assigned as a child concept of “Quan lai”.

In the other hand, because a dictionary is built by linguists, so sometimes the WordNet dictionary is not for decentralized accurate results from the course of history. This difficulty arises due to the intensive language learning is not necessarily in the field of history. To overcome this problem we have combined the reference ontology VHO reference to the historical ontology of projects [3, 4 5] to formalize the process of taxonomy extractor. Results of hierarchical concepts of the “historical character” concept can be shown below.

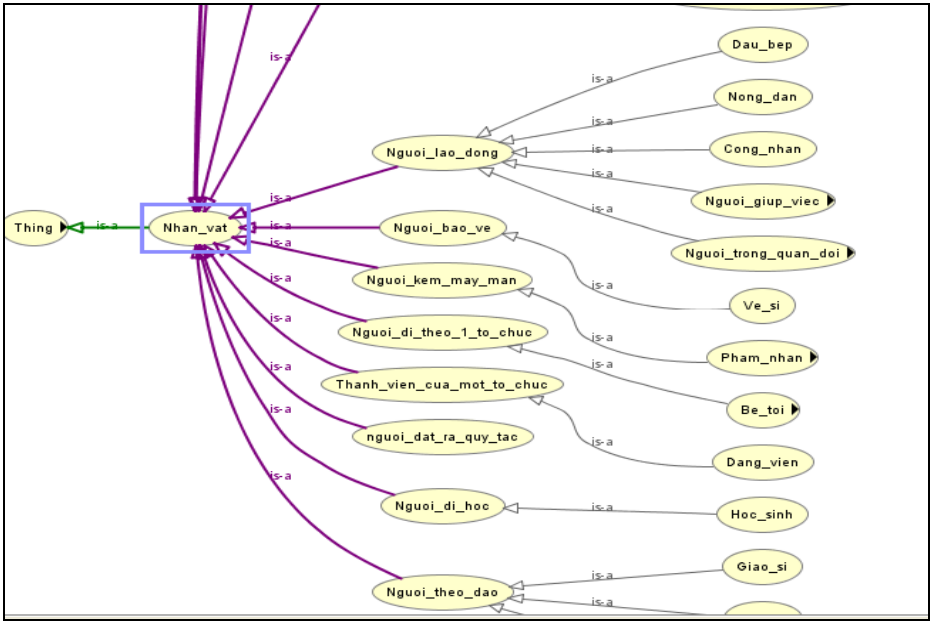


Fig. 1. results of hierarchical concepts of the "historical character" concept

We can summarize statistics result of ontology VHO in this table below:

Table 2. VHO ontology statistics

Concept or entities	**Number**
Concept	165
Kinds of "historical character"	51
Entities of "historical event"	63
Entities of "historical character"	21

2.3 Property Information Extraction

The key requirement of this step, is identify properties that a concept/ historical characters should have/ do. For example, with the "King" concept, we have properties such as "to be a leader" or "against the enemy"… Because of this, in term extraction process of properties of a concept, we will find out properties that appear near the concept with high frequency in history articles. For the example above, we have to find properties that "King" usually have. Some most common "King" action can be shown below.

Table 3. List of common properties of the “King“ concept

Properties (Vietnamese)	English meaning
Danh giac	Against the enemy
Len ngoi vua	Be a leader
Xu tu, chem dau	Execute (kill) someone who have serious crimes
Chi huy	Lead the army

2.4 Storage Change Entities of VHO Ontology

Based on the change entities storage model given in [2], we refer to overcome some shortcomings of this model to provide storage model of entity changes of VHO ontology. For an existing entity such as Vietnam, a nation, we collects set of entities that now they change to be Vietnam. They are Au Lac, Giao Chi, Dai Viet, North Vietnam or South Vietnam, a set of ancient nation of Vietnam in the past. These nation will be connected to each other by a events that make changes from one country to another country(s), called ChangeCauseEvent. Like other events, this kind of events happened in the timeline or in a time interval. Based on Figure 2, we can identify the Au Lac nation changes to Giao Chi by a war events (ChangeCauseEvent) that happened in 179 B.C. To be able to easily make semantic queries on the entity changes, we provide a mapping between the all past entities and a present entity through the object property, called “Has current entity”. For example: Au Lac “Has current entity” Vietnam. In this figure below, the “Has current entity” property is illustrated by a red arrow in the right of nation entities.

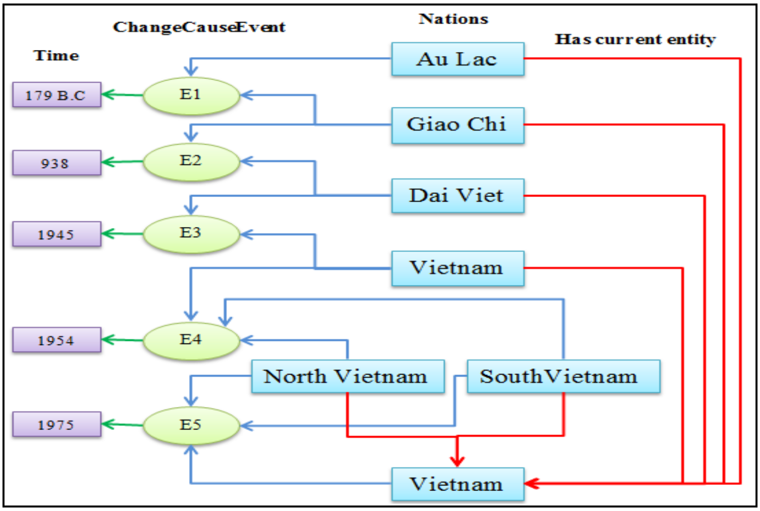


Fig. 2. Storage change entites model in VHO ontology

2.5 Storage Time Entities in VHO Ontology

A timeline or a time interval in VHO application is saved as an object of concept Time in the VHO ontology. There are some conventions to set name for entities of concept Time to help application to be able to obtain information about the type of time interval or timeline through the time analysis from ontology tools we have developed. We can take a few examples:

Table 4. Time entities in the VHO ontology

Time entity name	Time type	Mean
A_yy_1A1A1479BB0A0A0	timeline, only year	Year 1479
B_yy_1A1A1418BB0A0A0_yy_1A1A1423BB0A0A0	Time interval, only year in each timeline	Year 1418 to 1423

In addition to naming conventions of Time entities, time storage method in VHO still need to be assured of easy ability to sorting by time for historical objects through semantic queries. Obviously this is a very important field of history, when the query have to be arranged. For example, if we have a query "find the events that occurred from 1010 to 1100", query result have to sort all events in this period by time. To be come over sorting by time, each entity of Time concept must have 2 properties: "Start day of year" and “Start year”, and Time entities that reflects time interval must have 2 more properties, "End day of year" and “End year”. This storage method makes it easy to arrange history objects from time to time based on numerical parameters. For example, with the Time entity name A_ddmmyy_19051890BB0A0A0, we have “Start day of year” 139 (number 139 in 365 days of the year 1890) and “Start year” 1890. With conventional methods of storage Time entities, the VHO ontology have been able to meet two storage requirements about identifying time type and arranging historical objects in VHO application effectively.

3 System Architecture

The VHO application was built on the basis of semantic softwares have been used by a large user community. They are the ontology editor Protégé ([11]), the ontology storage server Allegrograph ([12]), and Jena API ([13]), a framework programming with semantic applications.

Through the Application layer, we will put into the system semantic question. Question will be converted to SPARQL 1.1 query by the Build SPARQL query package in the Semantic processing layer. Through the SPARQL query processing package in this layer, and the SPARQL query will access the ontology repository

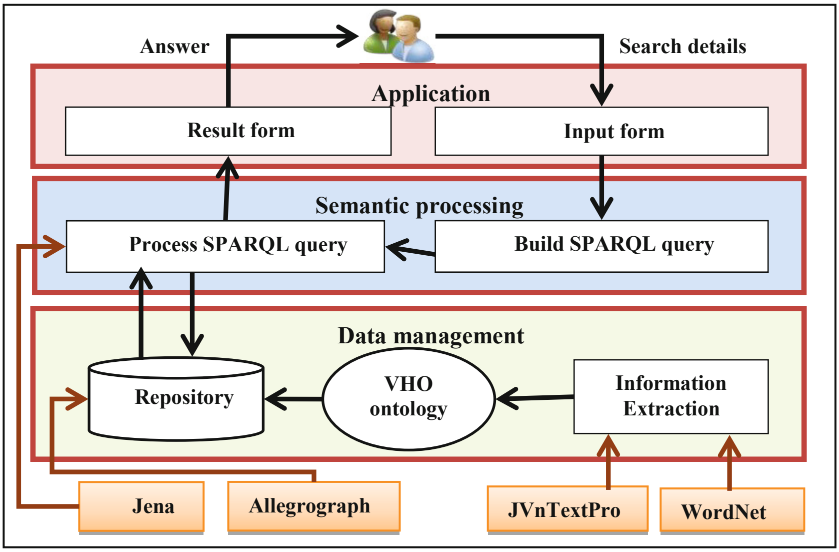


Fig. 3. The VHO system architecture

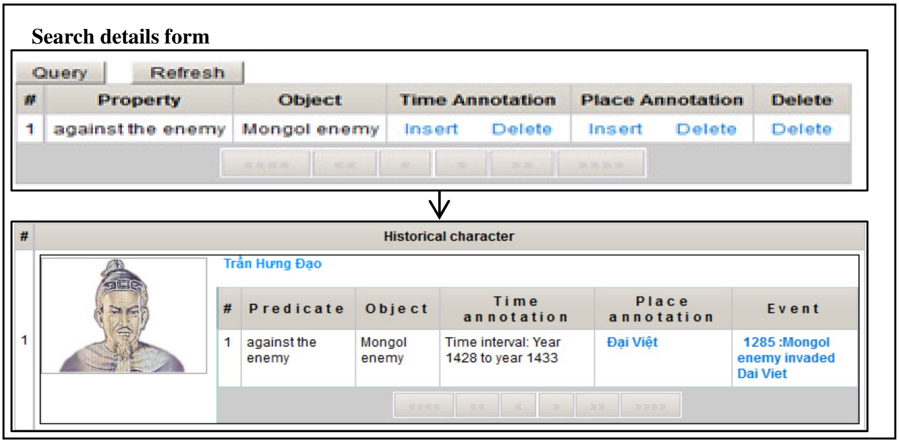
AllegroGraph, by the semantic programming framework Jena API. The ontology repository is located in the Data management layer. Queries results will be sent to the Application layer to answer user's question.

4 Experimental

4.1 Semantic Search Function of the VHO Application

Semantic search function in VHO allows users to search for natural questions, which are form of predicate-object question. The VHO application allows users to search with of six types of object: time, place, organization, character, event and types of string. With the semantic search function for historical characters, users will enter into detailed features/operation of historic character and be able to enter or not enter the corresponding object to make a predicate-object question. In addition, users can also add their annotation about the time or place, that is when and where the predicate appears. Along with common predicate-object questions, the VHO search function also be able answer complex historical questions related to time and change entities, which reflects specifics of historical domain. Address of the application along with the data related to the paper can be accessed at The VHO Introduction site[1].

[1] http://phandanghung.wordpress.com/2012/08/03/welcome-to-the-vho-project/



Search details form

#	Property	Object	Time Annotation		Place Annotation		Delete
1	against the enemy	Mongol enemy	Insert	Delete	Insert	Delete	Delete

#	Historical character
1	Trần Hưng Đạo

#	Predicate	Object	Time annotation	Place annotation	Event
1	against the enemy	Mongol enemy	Time interval: Year 1428 to year 1433	Đại Việt	1285 :Mongol enemy invaded Dai Viet

Fig. 4. Semantic search interfaces and details of a historical characters who adapt question 1.1 (Answer: Tran Hung Dao, General of Dai Viet nation)

4.2 Compare VHO with Keyword Search and Relational Database Search

4.2.1 Building Test Method

To evaluate the accuracy of VHO project compared to other search method in the field of history, we collect a set of 50 questions about finding historical charaters, which were obtained from the learning history website [16] to make a test data input. These questions are divided into three query types:

1. Query name of historical characters directly: Search by typing the historical characters directly. For example: "Le Loi", "Tran Hung Dao"…
2. Query by predicate-object question: Search by typing in features/ operations and object of characters. For example: question 1.1.
3. Search by referent predicate-object question: features/ operations and the concept of looking for a specific case, also known as inherited from class activities / characteristics and the search concept store in the test database. For example: question 1.2.

With keyword search, we will search by keyword on the website history [16] via the basic and advanced search functions of this website. With relational database search method, we store data of answers the question in a relational database. This database has 2 tables, which save information details and save predicate/ object of historical characters accordingly. Data about the answers of the 50 questions will be entered into the data repository of each method.

4.2.2 Test Results

Test results show that the search method using keyword has a relatively low accuracy with the test data. By this method, search function can only find information on historical characters if user enter the query type 1.

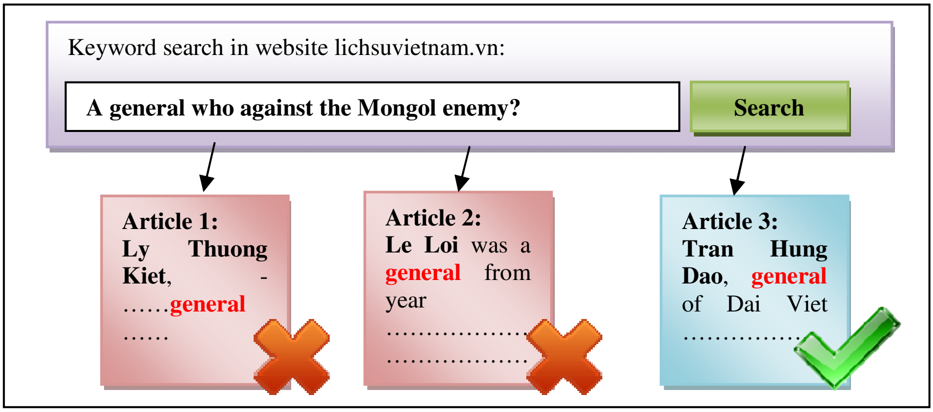

Keyword search in website lichsuvietnam.vn:
A general who against the Mongol enemy?
Search
Article 1: Ly Thuong Kiet, - ……general ……
Article 2: Le Loi was a general from year ……………… ………………
Article 3: Tran Hung Dao, general of Dai Viet ……………

Fig. 5. Incorrect anwser of question 1.1 by keyword search function in website [16]

For example, let's look at the answer of question 1.1 in website [16] keyword search function described in figure 5. Cause of the "general" keyword, website [16] provide a set of Vietnamese articles which have "general" keywords, they are articles about Vietnamese generals, include who didn't against the Mongol enemy. That causes incorrect answers. This limitation has been overcome when searching in database, which take a higher accuracy. However, the method searches the database does not give an correct answer when users input reference predicate-object questions (question type 3). The comparison of 3 search methods can be shown in figure 6.

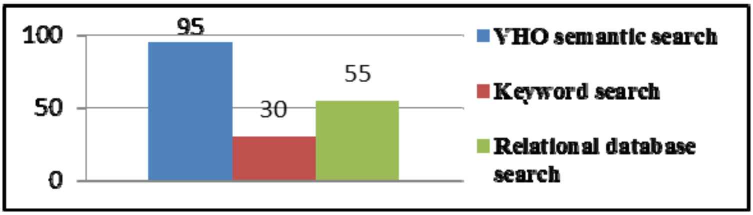


Fig. 6. The comparison chart by percentages of 3 search methods in 50 questions test

Based on the comparison chart above, we can clearly see effect of semantic search methods in VHO application with high accuracy (95%) compared to 2 other methods. This accuracy comes from the utility of semantic technology that allow answer natural questions in form of predicate-object question and reference predicate-object question.

5 Conclusion

Within the scope of this paper, we present the steps to build the semantic information search system for Vietnamese history VHO. In Vietnam, semantic search in the field of history is a relatively new issue. The research and experimental results of VHO system showed that this project initially gives users, especially history learners new ways to look up historical information easily and recall. Issues of ontology design for Vietnamese history can also be applied similarly to the field of the history of other countries, when we build the historical ontology based on the historical documents of those countries. In studies in future, we will make new researches to improve the quality of semantic search gain the absolute accuracy and develop VHO to be a smart auto question-answer system in the field of history.

Acknowledgements. We want to say a sincerely thank to the National Foundation for Science and Technology Development (NAFOSTED) Foundation of Vietnam Government who supports us to publish this research in the ACIIDS 2013 conference.

References

1. David Taniar, Johanna Wennt Rahayu: Web Semantics Ontology. IDEA Group Publishing (2006)
2. Ide, N.: David Woolner: Historical Ontology. In: Words and Intelligence II, January 01, 2007, pp. 137–152 (2007)
3. Clifford, J.: Natural Language Querying Of Historical Databases. Journal Computational Linguistics archive 14(4), 10–34 (1988)
4. Cao, T.-D., Phan, T.-H., Nguyen, A.-D.: An Ontology Based Approach to Data Representation and Information Search in Smart Tourist Guide System. In: 2011 Third International Conference on Knowledge and Systems Engineering, KSE, pp.171–175 (2011)
5. Katifori, A., Nikolaou, C., Platakis, M., Ioannidis, Y., Tympas, A., Koubarakis, M., Sarris, N., Tountopoulos, V., Tzoannos, E., Bykau, S., Kiyavitskaya, N., Tsinaraki, C., Velegrakis, Y.: The Papyrus Digital Library: Discovering History in the News. In: Gradmann, S., Borri, F., Meghini, C., Schuldt, H. (eds.) TPDL 2011. LNCS, vol. 6966, pp. 465–468. Springer, Heidelberg (2011)
6. Fahmi, I., Ellerman, H.: Project Proposal: SWHi, Applying the Semantic Web for History: Early American Imprints, Series I (2006), `http://www.citeulike.org/user/AlisonBabeu/article/1373371`
7. Nagypál, G.: History Ontology Building: The Technical View. In: The Proceedings of the XVI International Conference of the Association for History and Computing (AHC 2005), pp. 207-214. Royal Netherlands Academy of Arts and Sciences, Amsterdam (2005)
8. Noy, N.F., McGuinness, D.L.: Ontology Development 101: A Guide to Creating Your First Ontology. Stanford Knowledge Systems Laboratory Technical Report KSL-01-05 and Stanford Medical Informatics Technical Report SMI-2001-0880 (March 2001)
9. Nguyen, C.-T., Nguyen, T.-K., Phan, X.-H., Nguyen, L.-M., Ha, Q.-T.: Vietnamese Word Segmentation with CRFs and SVMs: An investigation. In: The 20th Pacific Asia Conference on Language, Information, and Computation (PACLIC), 1st-3rd, Wuhan, China (November 2006)
10. Wordnet, `http://wordnet.princeton.edu/`
11. Protégé, `http://protege.stanford.edu/`
12. Allegrograph, `http://www.franz.com/agraph/allegrograph/`
13. Jena API, `http://jena.apache.org/`
14. Ngo, S-L.: Complete Annals of Dai Viet (2001), `http://en.wikipedia.org/wiki/Đại_Việt_sử_ký_toàn_thư`
15. Pham, V.-S. : Vietnamese Complete History (1995), `http://vi.wikipedia.org/wiki/Việt_sử_toàn_thư`
16. Learning Vietnamese History website, `http://www.lichsuvietnam.vn/`
17. Best of History Sites website, `http://besthistorysites.net/`

Spam E-Mail Classification Based on the IFWB Algorithm

Chichang Jou

Department of Information Management, Tamkang University
Tamsui, New Taipei City, Taiwan 25137
cjou@mail.tku.edu.tw

Abstract. The problem of spam e-mails has been addressed for some time. Most of the solutions are based on spam e-mail classification and filtering. However, the content of spam e-mails drifts with new concepts or social events. Thus, several spam classifiers perform effectively when their models are initially established, and their performances deteriorate with time. A learning mechanism is required to adjust the classification parameters for new and old e-mails. Because of the spread of spam e-mails, the number of spam e-mails is larger than that of legitimate e-mails. Therefore, most classifiers produce high recall for spam e-mails and low recall for legitimate e-mails. Based on the Bayesian algorithm, we propose an incremental forgetting weighted algorithm with a misclassification cost mechanism that extracts features by IGICF (Information Gain and Inverse Class Frequency) to address the problem of concept drift and data skew in spam e-mail classification. We implemented the algorithm and performed detailed tests on the effectiveness of the mechanism.

Keywords: spam classification, incremental forgetting, misclassification cost.

1 Introduction

The proliferation of spam e-mails has caused considerable management costs for organizations and individuals. Several spam e-mail management mechanisms have been proposed, such as the use of legislation, the use of SenderIDs to enhance the SMTP protocol, the establishment of blacklists and white lists, and the use of machine learning mechanisms to filter spam. Among them, the use of a machine-learning mechanism to perform e-mail content and/or user behavior analysis is popular. Examples include Bayesian (Sahami et al., 1998; Hayat et al., 2010; Almeida et al., 2011), Support Vector Machine (SVM) (Drucker et al., 1999; Sculley & Wachman, 2007; Tseng & Chen, 2009), and KNN (Delany et al., 2005; Alguliev et al., 2011).

In spam classification, with changes of user preferences and updates of spam characteristics by spammers, new spam topics, content, and characteristics frequently emerge. Previous spam solutions usually performed effectively in the beginning. If they do not adjust their model parameters, their effectiveness gradually declines. Thus, it is crucial to manage concept drift (Widmer & Kubat, 1996) in spam classification.

A. Selamat et al. (Eds.): ACIIDS 2013, Part I, LNAI 7802, pp. 314–324, 2013.

Symantec reported that in 2010, the average global spam rate for the year was 89.1%, with an increase of 1.4% compared with 2009. Under such a data distribution imbalance, most spam classifiers perform effective classification for spam e-mails, but exhibit inferior performance with a low recall for legitimate e-mails (Chawla et al., 2004). Mechanisms to maintain an effective recall for legitimate e-mails are required.

Based on Bayesian classification, we propose an Incremental Forgetting Weighted Bayesian (IFWB) algorithm with a misclassification cost mechanism. The IFWB integrates IGICF (Information Gain and Inverse Class Frequency) (Xu et al., 2007) to extract textual features for e-mails, incremental forgetting mechanism (Koychev, 2000), and various misclassification cost assignments (Fawcett, 2004) to solve the problems of concept drift and data skew. We implemented IFWB and tested the feasibility of IFWB using well-studied e-mail data sets.

The rest of the paper is organized as follows: Section 2 presents related work; Section 3 introduces the IFWB algorithm; Section 4 provides the experiment details and their results; and lastly, Section 5 offers a conclusion and future research directions.

2 Related Work

Based on KNN, Delany et al. (2005) built association networks for cases and features, and deleted noise and duplicate cases in the instance e-mail set. When a misclassification occurred, they inserted the new case into the correct class and deleted old cases for that class.

Fdez-Riverola et al. (2007) combined Bayesian classification, relevant term identification (RTI), and representative message selection (RMS) to perform e-mail classification. Through RTI, the most descriptive words for an incoming e-mail were selected. The e-mail was subsequently classified based on the statistics of these words. RMS regulates the number of added and deleted cases each time.

Tseng & Chen (2009) proposed an incremental SVM model for spam detection based on e-mail user behavior. Features of each user were extracted to train the SVM model. Moreover, to catch the evolving nature of e-mail communication, they presented an incremental update scheme to efficiently re-train the model.

Hayat at al. (2010) proposed an adaptive spam filtering system based on language model. It would detect concept drift based on computing the deviation in email contents distribution.

Almeida et al. (2011) examined how the term-selection techniques affect the categorization accuracy of different filters based on the Bayesian decision theory.

Alguliev et al. (2011) developed a genetic algorithm for solving spam clustering problems, where the objective function was a maximization of similarity between messages in clusters, which was defined by a KNN algorithm.

3 THE IFWB Algorithm

3.1 Choices of Related Modules

Among the spam classifications, the naïve Bayesian algorithm exhibits excellent classification performance and requires less execution time. Thus, we selected the naive Bayesian algorithm as our base algorithm.

Prior studies revealed two main approaches to manage concept drift of spam e-mails, as follows: (1) case-based (Widmer & Kubat, 1996; Koychev, 2000): this approach adjusts the classification framework using case selection, assigning weights to cases, or using ensemble results of several methods; and (2) feature set update (Delany et al., 2005; Fdez-Riverola et al., 2007; Tseng & Chen, 2009): this approach periodically updates features of e-mails to maintain good classification power.

In case-based management of classification, if old cases are directly removed from the data set, then some crucial information contained in these cases disappears, which may affect the classification. Instead of ignoring the old cases, we adjusted their weight of importance. We used the moderate gradual forgetting mechanism proposed by Koychev (2000). For feature set update, we used the incremental feature selection method proposed by Katakis et al. (2005).

For timing of the feature set update, previous studies performed a novel feature set update either for each incoming case (Fdez-Riverola et al., 2007; Katakis et al., 2005), or regularly for a fixed duration (Delany et al., 2005). We adopted a moderate frequency such that an update would be performed only when a misclassification occurs.

According to Monard & Batista (2002), three methods to manage data skew are as follows: (1) majority reduction; (2) minority increase; and (3) assigning misclassification cost. Majority reduction tends to erase crucial information in the data set, and a minority increase may induce noise. Following Fawcett (2004), we assigned misclassification costs to tackle the inherent data skew problem.

3.2 Flow Charts of IFWB

Two phases are used in IFWB: training and classification. Figure 1 shows the four steps in the training phase, as follows: data preprocessing, dictionary buildup, feature selection, and vector model build up. These steps are explained as follows:

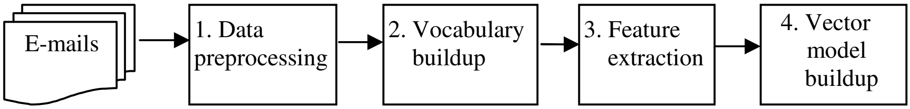


Fig. 1. Flow chart of IFWB's training phase

1. Data preprocessing: we preserved the subject and body of each e-mail. We subsequently deleted the punctuations and HTML tags. For an English data set, we used blanks as the separator for tokenization. For a Chinese data set, we used the MMSEG tool to extract Chinese terms. Subsequently, we used the stop words in the bow toolset developed by McCallum to filter out non-feature words. Finally, we used the Porter stemming method (Porter, 1980) to transform each word into their root formats.
2. Vocabulary buildup: to avoid recalculating the term frequency statistics for previous cases, we used the incremental feature extraction method through term vocabulary buildup (Katakis et al., 2005). We stored the term frequency statistics for the terms in the vocabulary. These statistics would be used to extract the feature set when it requires updating.

3. Feature extraction: From the statistics information in the vocabulary, we used the IGICF algorithm (Xu et al., 2007) to extract keywords with top IGICF values. Details of IGICF are provided in Section 3.3.
4. Vector Model Buildup: Suppose there are n features. Each e-mail X is subsequently represented as a Boolean vector: $X = (x_1, x_2, ..., x_n)$. In this vector, if the *i*th feature appears in X, $x_i = 1$; otherwise, $x_i = 0$.

Figure 2 shows the flowchart of the classification phase for an incoming e-mail. Details of the classification are as follows:

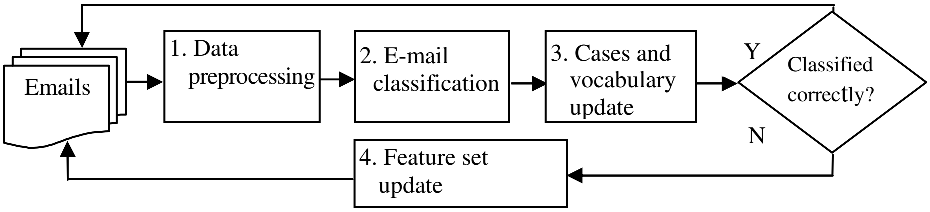


Fig. 2. Flow chart of the classification phase of IFWB

1. Data preprocessing: This step is similar to that in the training phase. The statistics of term frequencies for an e-mail were calculated. The e-mail was represented as a vector according to the current feature set.
2. E-mail classification: We used the proposed IFWB, which combines Bayesian classification with a linear gradual forgetting mechanism and assigned misclassification cost. Details of the classification and cost assignment are presented in Sections 3.4 and 3.5.
3. Cases and vocabulary update: After e-mail classification, the e-mail is collected to its correct class, and its term statistics are updated to the vocabulary.
4. Feature set update: We updated the feature set according to IGICF only when a misclassification occurred.

3.3 Feature Set Extraction

Algorithms for feature set extraction include document frequency, information gain, and chi-square statistics. Yang et al. (1997) performed experiments to compare the performances of these algorithms, and found that the algorithm based on information gain produced feature sets with optimal classification results. The equation to compute the information gain for a term is as follows:

$$IG(t) = P(t)\sum_{i=1}^{m} P(c_i \mid t)\log P(c_i \mid t) + P(\bar{t})\sum_{i=1}^{m} P(c_i \mid \bar{t})\log P(c_i \mid \bar{t}) - \sum_{i=1}^{m} P(c_i)\log P(c_i), \quad (1)$$

where *t* represents the event in which term t appears, $\bar{t}$ represents the event in which term t does not appear, and m is the number of classes in the cases.

Based on the information gain algorithm, Xu et al. (2007) considered the inverse class frequency (ICF). When the class frequency of a term is low, the term has a higher distinguishing power. Their experiments showed that the IGICF algorithm can extract a feature set with a superior classification result. The equation for IGICF is as follows:

$$ICF(t) = \log\frac{m}{m_t} + 1 \qquad \text{and} \qquad IGICF(t) = IG(t) * ICF(t), \tag{2}$$

where m is the number of classes in the case, and m_t represents the number of classes containing term t. The "+1" in the equation for ICF(t) in (2) is used to avoid a value of 0 when t appears in every class.

3.4 Weighted Bayesian Classification

We assume that the importance of an e-mail gradually decreases with time. Conversely, for each class, a more recent e-mail is more important. For example, the most recent spam e-mail is more important than the recent 100th spam e-mail. Therefore we defined a weighting function *w=f(t)* to compute the weight of recency for an e-mail at time t; t was represented by the order of its recency. Figure 3 shows a linear gradual forgetting function for the weighting function. From this gradual forgetting function, we computed weight of the most recent *i*th case in each class w_i, as follows:

$$w_i = -\frac{2k}{N-1}(i-1)+1+k \qquad \text{and} \qquad \frac{\sum_{i=1}^{N} w_i}{N} = 1, \tag{3}$$

where k is the forgetting speed (the value of k is between 0 and 1), and N is the number of cases in the class.

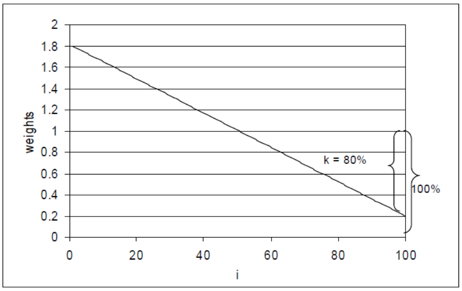


Fig. 3. Gradual Forgetting Function with N=100 and k=80% (Koychev, 2000)

Using the gradual forgetting function, we computed the weight of recency of each case in its class. We subsequently applied the weight in naive Bayesian computing, as follows:

$$P(c_j \mid X) = \frac{P(X \mid c_j)P(c_j)}{P(X)} \quad \text{and} \quad P(X \mid c_j) = \prod_{q=1}^{n} P(x_q \mid c_j) = \prod_{q=1}^{n} \frac{\sum_{i=1}^{N_j} w_{ij} * \theta_{x_q = i_{j,q}}}{N_j} \tag{4}$$

Suppose e-mail X is represented by n features, and x_q is the qth feature value. Let N_j be the number of e-mails in class $i_{j,\, q}$ be the qth feature value of the most recent ith e-mail in class c_j, and w_{ij} is the weight of recency for the ith e-mail in class c_j. If $x_q = i_{j,q}$, $\theta_{x_q=i_{j,q}} = 1$; otherwise, $\theta_{x_q=i_{j,q}} = 0$.

If w_{ij} is not included in equation (4), equation (4) represents the traditional naïve Bayesian. Considering the weight of importance, suppose $x_q=1$; subsequently, $P(x_q = 1 \mid c_j)$ represents the sum of all weights of e-mails in class c_j that satisfy the condition in which the qth feature value is 1, divided by the number of e-mails in class c_j. When the weight of importance follows the gradual forgetting function, a more recent e-mail is more important. This computation can be used to make a spam decision considering the recency of e-mails.

To avoid a result in which $P(x_q \mid c_j)$ has a value of 0 for some q, which causes $P(X \mid c_j)$ to have a value of 0 regardless of the performance of the other features, we used Laplacian correction to provide a small value for this case. The modified equation for the naïve Bayesian is as follows:

$$P(x_q \mid c_j) = \frac{(\sum_{i=1}^{N_j} \theta_{x_q=i_{j,q}}) + 1}{N_j + 2} \tag{5}$$

By considering the weight of recency, this equation may produce a medium value. Therefore, we modified the equation as follows:

$$P(x_q \mid c_j) = \frac{(\sum_{i=1}^{N_j} \theta_{x_q=i_{j,q}} + 1) * w_{N_j j}}{N_j + 2}, \tag{6}$$

where $w_{N_j j}$ is the smallest weight of all cases in class c_j.

3.5 Misclassification Cost

In assigning the misclassification cost, we defined the misclassification cost matrix in Table 1, where c_{00}, c_{01}, c_{10}, and c_{11} are the costs for classifying spam/legitimate e-mails as spam/legitimate e-mails. Because c_{00} and c_{11} are the costs for cases that are classified correctly, their values are 0.

Table 1. Misclassification cost matrix

predicted \ actual	spam	legitimate
spam	c_{00}	c_{01}
legitimate	c_{10}	c_{11}

Considering misclassification cost, the probabilistic cost function of spam/legitimate e-mails (PCF_{spam} / PCF_{legit}) are represented as follows:

$$PCF_{spam}(X) = \frac{P(spam | X) * c_{10}}{P(spam | X) * c_{10} + P(legit | X) * c_{01}} \text{ and } PCF_{legit}(X) = \frac{P(legit | X) * c_{01}}{P(spam | X) * c_{10} + P(legit | X) * c_{01}} \quad (7)$$

where $P(spam | X)$ and $P(legit | X)$ are the weighted probability obtained in Section 3.4. Therefore, our classification decision was determined using the two values in these equations. For data skew, the number of spam cases was larger than that of legitimate cases. Therefore, the recall rate of legitimate e-mails was inferior. A higher misclassification cost can be assigned to legitimate e-mails to increase the recall rate.

4 Experiments

We tested the IFWB algorithm on three public e-mail corpuses, as follows: LingSpam, Spamassassin, and TREC (Spam Track Corpus). LingSpam has been tested in several previous studies. The e-mail data in Spamassassin include the header and body. TREC is in Simplified Chinese in the GB2312 codes. For LingSpam, we extracted 962 e-mails, 50% of which were spam e-mails. For Spamassassin, we filtered out e-mails with special encoding and obtained 900 cases, 50% of which were spam e-mails. We extracted 1800 e-mails for the TREC cases.

In the first experiment, we evaluated the performance of IFWB in these corpuses. In the second experiment, we tested the effect of ICF by comparing the performances of the IGICF and IG algorithms. In the third experiment, we tested the performances of the naïve Bayesian and IFWB for cases under concept drift. In the fourth experiment, we tested the effect of the forgetting speed on the classification performance. In the fifth experiment, we investigated the effect of misclassification cost on various degrees of data skew. In the sixth experiment, we assessed the effect on performance under various misclassification costs.

The forgetting speed in Experiments 1, 2, 3, 5, and 6 were fixed at 80%. The misclassification cost for legitimate e-mails in Experiments 1, 2, 3, and 4 was fixed at 2.5 for LingSpam, 3.5 for Spamassassin, and 1 for TREC. The number of features for LingSpam and Spamassassin was set to 120. Because several cases in TREC had similar e-mail content, the number of features for TREC was set to 300 to increase the recognition rate for small classes. We used classification accuracy, spam recall, and legitimate recall as our main performance indicators, as follows:

$$accuracy = \frac{n_{s \to s} + n_{l \to l}}{n_{s \to s} + n_{s \to l} + n_{l \to l} + n_{l \to s}} \quad (8)$$

$$spam\_recall = \frac{n_{s \to s}}{n_{s \to s} + n_{s \to l}} \quad (9)$$

$$legitimate\_recall = \frac{n_{l \to l}}{n_{l \to l} + n_{l \to s}}, \tag{10}$$

where $n_{s \to s}$, $n_{s \to l}$, $n_{l \to l}$, and $n_{l \to s}$ are the number of cases in which spam and legitimate e-mails were classified as spam and legitimate e-mails.

Figure 4 shows the classification results for the three corpuses in the first experiment. As shown in the figure, Acc represents accuracy, SR represents spam recall, and LR represents legitimate recall. The accuracies for the three corpuses were approximately 97%-98%. The result of the Spamassassin corpus was superior to the 94%-95% result obtained by ECUE. For the TREC corpus, the legitimate recall was 100%, whereas the spam recall was 94%. This occurred because of the large number of spam e-mails with similar subjects and almost the same body. When a spam e-mail does not contain feature words of those similar subjects, the e-mail is misclassified as legitimate.

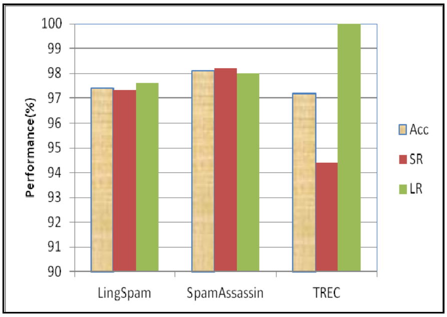


Fig. 4. Classification results of IFWB for three corpuses

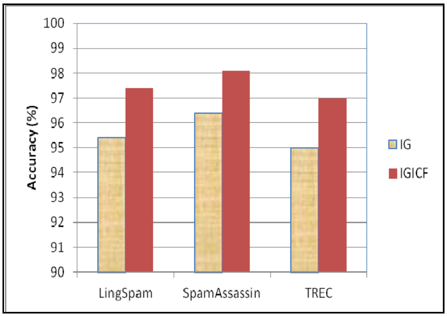


Fig. 5. Accuracy of IG and IGICF feature extraction methods for three corpuses

In the second experiment, we tested the accuracy of the three corpuses when the feature set was extracted under the IG and IGICF algorithms. The results are shown in Fig. 5. As shown in the figure, the performance of IGICF is superior to that of IG.

In the third experiment, we used the Spamassassin corpus to test the performance of the naïve Bayesian and IFWB in managing the concept drift problem. We sorted the e-mails according to the date and time information in the header. A total of 100 e-mails were included in the training data set. In the testing data set, we used 100, 300, 900, and 1500 e-mails that were close to the training data set in time. The results are shown in Fig. 6. High classification accuracy can be maintained because IFWB adjusts the classification framework for new incoming e-mails. However, the accuracy of the naïve Bayesian gradually decays as the concept drifts away.

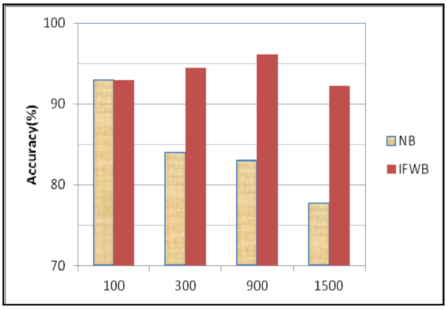


Fig. 6. Accuracy of Bayesian and IFWB under various concept drift sizes

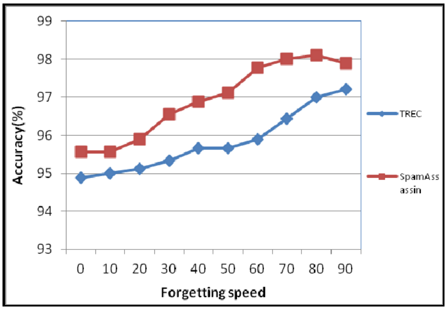


Fig. 7. Accuracy for different forgetting speeds

In the fourth experiment, we used Spamassassin and TREC corpuses to test the effect of forgetting speed. Because e-mails in Lingspam do not contain date and time information, it was not used in this experiment. We sorted the e-mails according to their date and time. The forgetting speed ranged from 0% to 90% in increments of 10%. The results are shown in Fig. 7. All cases with positive gradual forgetting exhibited superior performances to the case with a forgetting speed of 0%. In the TREC corpus, the optimal performance was achieved when the forgetting speed was 90%, whereas in the Spamassassin corpus, a forgetting speed of 80% achieved optimal performance. This may be caused by the neglect of some important information when the forgetting speed is excessively high.

In the fifth experiment, we tested the effect of the misclassification cost using the LingSpam corpus with various data skew degrees. In the training data set, we used 331 spam e-mails and 33, 66, 99, and 150 legitimate e-mails separately. We compared the legitimate recall performance for the cases with and without considering the misclassification cost. The results are shown in Fig. 8. As shown in the figure, when the data skew degree is high, the performance for the case without considering the misclassification cost is relatively inferior. For the case considering the misclassification cost, the performances are excellent.

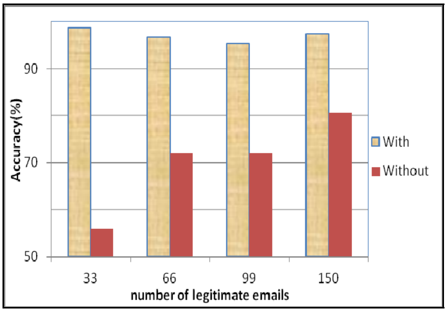


Fig. 8. Performance comparisons with and without misclassification cost consideration for various data skew degrees

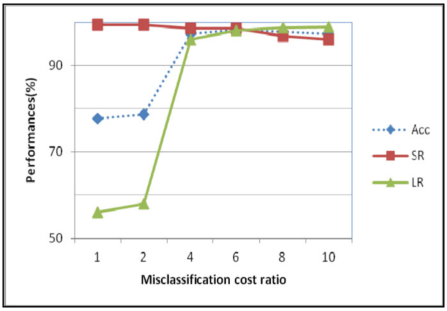


Fig. 9. Performance comparisons for various misclassification cost ratios

In the sixth experiment, we used the LingSpam corpus with a fixed data skew of 331 spam e-mails and 33 legitimate e-mails to test the effect of various misclassification costs. The misclassification cost of legitimate e-mails was set to 1, 2, 4, 6, 8, and 10 times that of spam e-mails. Figure 9 shows the three performance indicators for these cases. As shown in the figure, when the cost ratio is in the range of 2 to 6, the legitimate recall rates rise in conjunction with the cost ratio. When the cost ratio is larger than 6, the legitimate recall rates increase slowly, and the accuracy decreases slightly. Thus, a high cost ratio may not be preferable. The proper cost ratio must be decided based on the data skew degree.

5 Conclusion

Because concepts in e-mails drift, e-mail classification accuracy decreases with time. Additionally, when the number of spam e-mails is larger than that of legitimate e-mails, the recall of the legitimate e-mails is inferior in most e-mail classifiers. Based on the Bayesian algorithm, we combined the gradual forgetting and misclassification cost assignment mechanisms to address the problems of concept drift and data skew in spam e-mail classification.

In managing concept drift, the introduction of new e-mails affects the classification framework. However, it may also introduce noise to the framework. The classifier can easily make an incorrect decision for the cases with lower numbers, especially when several cases in a class have similar contents. Thus, we will incorporate other mechanisms to manage the noise and duplicate cases to improve our algorithm.

References

1. Alguliev, R.M., Aliguliyev, R.M., Nazirova, S.A.: Classification of textual e-Mail spam using data mining techniques. Applied Computational Intelligence and Soft Computing, Article ID: 416308 (2011)
2. Almeida, T., Almeida, J., Yamakami, A.: Spam filtering: how the dimensionality reduction affects the accuracy of Naive Bayes classifiers. Journal of Internet Services and Applications 1(3), 183–200 (2011)
3. Chawla, N.V., Japkowicz, N., Kotcz, A.: Editorial: special issue on learning from imbalanced data sets. ACM SIGKDD Explorations Newsletter 6(1), 1–6 (2004)
4. Delany, S.J., Cunningham, P., Tsymbal, A., Coyle, L.: A case-based technique for tacking concept drift in spam filtering. Knowledge-Based Systems 18(4-5), 187–195 (2005)
5. Drucker, H., Wu, D., Vapnik, V.N.: Support Vector Machines for Spam Categorization. IEEE Transactions on Neural Networks 10(5), 1048–1054 (1999)
6. Fawcett, T.: In vivo spam filtering: a challenge problem for data mining. ACM SIGKDD Explorations Newsletter 5(2), 140–148 (2004)
7. Fdez-Riverola, F., Iglesias, E.L., Díaz, F., Méndez, J.R., Corchado, J.M.: Applying lazy learning algorithms to tackle concept drift in spam filtering. Expert Systems with Applications 33(1), 36–48 (2007)

8. Hayat, M.Z., Basiri, J., Seyedhossein, L., Shakery, A.: Content-Based Concept Drift Detection for Email Spam Filtering. In: 5th International Symposium on Telecommunications, pp. 531–536 (2010)
9. Katakis, I., Tsoumakas, G., Vlahavas, I.: On the Utility of Incremental Feature Selection for the Classification of Textual Data Streams. In: Bozanis, P., Houstis, E.N. (eds.) PCI 2005. LNCS, vol. 3746, pp. 338–348. Springer, Heidelberg (2005)
10. Koychev, I.: Gradual Forgetting for Adaptation to Concept Drift. In: Proceedings of ECAI 2000 Workshop Current Issues in Spatio-Temporal Reasoning, pp. 101–106 (2000)
11. Monard, M.C., Batista, G.: Learning with skewed class distributions. Advances in Logic, Artificial Intelligence and Robotics, 173–180 (2002)
12. Porter, M.F.: An algorithm for suffix stripping. Program (Automated Library and Information Systems) 4(3), 130–137 (1980)
13. Sculley, D., Wachman, G.M.: Relaxed Online SVMs for Spam Filtering. In: SIGIR 2007, pp. 415–422 (2007)
14. Sahami, M., Dumais, S., Heckerman, D., Horvitz, E.: A Bayesian Approach to Filtering Junk E-mail. In: Proceedings of the AAAI 1998 Workshop on Learning for Text Categorization, pp. 55–62 (1998)
15. Tseng, C.Y., Chen, M.S.: Incremental SVM model for spam detection on dynamic email social networks. In: Proceedings of CSE 2009 International Conference on Computer Science and Engineering, pp. 128–135 (2009)
16. Widmer, G., Kubat, M.: Learning in the presence of concept drift and hidden contexts. Machine Learning 23(1), 69–101 (1996)
17. Xu, Y., Li, J., Wang, B., Sun, C., Zhang, S.: A Study of feature selection for text categorization on imbalanced data. Journal of Computer Research and Development 44, 58–62 (2007) (In Simplified Chinese)
18. Yang, Y., Pedersen, J.O.: A comparative study on feature selection in text categorization. In: Proceedings of 14th Conference on Machine Learning, ICML 1997, pp. 412–420 (1997)

Meta Search Models for Online Forum Thread Retrieval

Research in Progress

Ameer Tawfik Albaham and Naomie Salim

Faculty of Computer Science and Information System,
Universiti Teknologi Malaysia, Skudai, Johor, Malaysia
ameer.tawfik@gmail.com, naomie@utm.my

Abstract. Online forum thread retrieval is the task of retrieving threads that satisfy a user information need. Several thread representations have been proposed, and it has been found that combining these representations outperformed the retrieval using the individual representations. However, these combining methods leverage query relevance judgments to rank threads. Furthermore, in online forums, obtaining relevance judgments is not an option. As a result, in this paper, we propose to combine various thread representations using meta search techniques; many meta search techniques do not require training and has been found to produce a competitive result to the approaches that use relevance judgments. Our experimental result shows two things. First, combining thread representations using meta search techniques is an effective approach. Second, the CombSUM or the CombMNZ meta search techniques outperformed the best baseline method on high precision searches.

Keywords: Forum thread search, Meta search.

1 Introduction

Online forums contain rich human generated knowledge. The knowledge is generated in the form of threaded discussion(threads). Each discussion consists of several text units known as the messages. A message is either an initial message or a reply message. An initial message initiates a discussion, and a reply message is a response to the initial message or to another reply in a thread. Although the forum content is available on the Web, accessing it is not an easy task due to the information overload problem. Thread search is one way to tackle that problem. However, traditional retrieval methods are unsuitable for thread retrieval [4, 9]. In a traditional retrieval method, the unit of retrieval is also the text unit. In contrast, in thread retrieval, the messages are the text units, whereas the threads are the retrieval units.

One method to solve the mismatch problem is to concatenate the thread message texts into a single virtual document and rank threads based on their virtual document relevance [4, 9, 2]. Another method is to represent a thread as a collection of messages and estimate the thread relevance by aggregating its

A. Selamat et al. (Eds.): ACIIDS 2013, Part I, LNAI 7802, pp. 325–334, 2013.

message relevance scores[4, 9]. In addition, an interpolation between these two representations' relevance estimates outperformed the individual representations [9]. [2] defined a thread as a structured document consists of three types of contents: the title content, the initial message content and the reply message content. It was found that combining these content types is better than retrieval using only a single content type.

In one hand, the experimental result of [9] and [2] supports combining various thread representations on ranking threads. On the other hand, both studies proposed methods that require relevance judgments. Constructing relevance judgments and training ranking functions is not an effective approach to thread retrieval for two reasons. First, it is known that constructing relevance judgments is a resource consuming activity. Second, in contrast to commercial web search companies that have the resource and the expertise to develop ranking models, online forums are community based platforms whose users and administrators have little knowledge about search algorithm design. Therefore, retrieval methods that do not require relevance judgments and prior knowledge in information retrieval will be more suitable to thread retrieval.

In information retrieval, the term "meta search" has been associated with the techniques that fuse several ranked lists of documents [12]. These ranked lists are generated by various retrieval methods. [6] approached the problem of searching using multiple document representations as a meta search problem. For each document representation, a list of documents was generated. Then, these lists were fused using meta search techniques such as CombSUM [10] and BordaFuse[1]. It was reported that the meta search approach had a competitive performance to that of the supervised methods on the known item search task [6]. Recently, [11] reported similar findings regarding the effectiveness of meta search techniques on two search tasks: homepage finding and topic distillation. Note that [6] and [11] addressed tasks that are not similar to thread ad hoc retrieval. In addition, [6] and [11]'s focus is finding web pages whose structure differs semantically and physically from that of threads[2, 9]. Nevertheless, their experimental results support the use of meta search techniques to search using multiple document representations.

In this paper, we report a part of an ongoing work to utilize meta search techniques to rank threads using multiple representations. The rest of this paper is as follows. Section 2 reviews related work to thread search and meta search techniques. Section 3 presents the used thread search methods and meta search techniques. The experimental design and result are given in Sections 4 and 5 respectively. Section 6 reports our discussion on the experimental result, and Section 7 concludes this paper and outlines future works.

2 Related Work

One challenge of thread retrieval is that the unit of retrieval is not the same as the text unit. In thread retrieval, the text units are the messages, whereas a user expects a ranked list of threads. [4] proposed two models to tackle this problem.

The first creates a virtual document for each thread by concatenating the thread message texts, then it scores threads based on their virtual document relevance to the query. In contrast, the second model defines a thread as a collection of text units (messages). Then, it scores threads by aggregating their message relevance scores. Several aggregation methods were used, and the best among them was based on the Pseudo Cluster Selection(PCS) method [8]. PCS scores threads in two steps: it scores a list of messages, then it ranks threads by taking the geometric mean of the top k ranked messages' scores from each thread. If a thread has less than k messages, PCS add a padding step with the minimum scoring message.

Nevertheless, PCS was not only applied to fuse message relevance scores, but also to fuse relevance scores of other types of text units. For instance, [9] treats a thread as a collection of several local contexts: posts, pairs, dialogues and the entire thread. The thread and the post contexts are identical to [4]'s virtual document and message based representations. In the pair and the dialogue contexts, the conversational relationship between messages is exploited to build text units. [9] estimates the relevance between the user query and a list of local "contexts"— the contexts could be messages, pairs or dialogues, then threads were scored by applying PCS to the threads' ranked contexts' relevance scores. [9] reported that the non linear interpolation between the thread context and the message, the pair or the dialogue context outperformed all retrieval methods based on the individual contexts.

Bhatia and Mitra[2] defined a thread as structured document: a document that consists of three small text units, which are the title, the initial message and the reply messages set. Then, the query term frequencies were given different weights based on where the term appears. Generally, [9] and [2]'s models are examples of retrieval using multiple representations.

Ogilvie and Callan[6] and Wu et al.[11] identified two approaches to retrieval using multiple representations: within search and after search. To estimate the relevance of a document, the within search based methods use a supervised approach on fusing the relevance scores coming from each representation. In other words, the within search methods require training to work effectively. The after search methods, known as the meta search or the data fusion approach, fuse ranked lists of documents that were generated by retrieval models based on the individual representations. It was found that the within search methods and the after search methods provide competitive performance on the finding known item[6, 11] and the topic distillation search tasks[11]. As a result, in this paper, we adopt the after search methods because many data fusions do not require training. In addition, considering that the thousands of forums that exist on forums vary in their abilities to construct training models, the meta search approach to thread retrieval suits such constraints better.

Based on the criteria used by a meta search technique, meta search techniques can be categorized into: score based and rank based fusion. The score based techniques — such as [10]'s CombSUM, use the relevance scores of documents, whereas the rank based techniques, [1, 5], utilize the ranking positions of these documents on the ranked lists.

In this paper, we restrict our discussion to the score based meta search methods. Some representatives of the score based techniques are [10]'s CombMIN, CombMAX, CombMED, CombANZ, CombSUM and CombMNZ. The first four methods score a document D using the minimum, maximum , median, arithmetic mean and sum of the relevance scores respectively. The CombMNZ method scores *D* by multiplying the sum of the relevance scores with the number of systems that retrieved *D*.

3 Fusion of Thread Search Methods

Based on our review on Section 2, there are seven different representations of a thread. Therefore, there are seven retrieval methods that can be used as an input for a meta search technique. The first four methods are based on [9]'s local contexts: thread, message, pair and dialogue contexts. The remaining retrieval methods are based on [2]'s structural components: title, initial message and reply messages.

However, in this paper, we focus our discussion on three methods: retrieval using the thread initial message, retrieval using the thread reply messages and retrieval using all types of messages. In all of them, we score a list of messages with respect to the user query, then we rank threads using these ranked messages' relevance scores.

The relevance between a user query $Q = \{q_1, q_2, \ldots, q_{|Q|}\}$ and a message m is estimated using the query language model[7] assuming term independence, uniform probability distribution for m and Dirichlet smoothing as follows[13]:

$$P(Q|m) = \prod_{i=1}^{|Q|} \left(\frac{tf(q_i; m) + \mu \frac{tf(q_i;C)}{|C|}}{|m| + \mu} \right) \tag{1}$$

where q_i is the i^{th} query term, μ is a free parameter and C denotes the message corpus. $tf(q;m)$ and $tf(q;C)$ are the term frequencies of q_i in m and C respectively. Lastly, $|m|$ is the message m's length, and $|C|$ is the corpus total number of tokens.

Ranking a thread T using its initial message $initial(T)$ is as follows:

$$\mathrm{I}(Q,T) = P(Q|initial(T)) \tag{2}$$

where $P(Q|initial(T))$ is calculated using equation 1.

Ranking threads using the reply message method(R) or the free message type method (M) requires an aggregation method to fuse the ranked messages' relevance scores. In [3], it was found the Pseudo Cluster Selection(PCS) method produced the best result. Therefore, this method is applied to the reply message and the free message type methods. However, instead of using the geometric of the scores, we used the sum of the scores. We tried using the sum and the geometric mean of the scores, but that the sum aggregation method was superior to the geometric mean. Let R_Q denote a list of ranked messages, R_T denote the set of all ranked messages from a thread T, and $R_{T,k}$ denote the set of the top k ranked messages. If the size of R_T is less than k, then an artificial ranked message R_{min} is padded to the thread T's list of ranked messages. Following this strategy, ranking threads using the free type messages (M) or the reply messages (R) is as follows:

$$PCS(Q,T) = \begin{cases} (k - |R_T|) \times R_{min} + \sum_{m \in R_T} P(Q|m) & \text{if } |R_T| < k \\ \sum_{m \in R_{T,k}} P(Q|m) & otherwise \end{cases} \tag{3}$$

where R$_{min}$ = $min_{m \in R_Q} P(Q|m)$ and $P(Q|m)$ is calculated using equation 1.

To generate the final list of threads, we used [10]'s score based methods: CombMIN, CombMAX, CombMED, CombANZ, CombSUM and CombMNZ. Note that all thread search methods are based on the language model framework [7, 13], hence we do not need to apply score normalization.

4 Experimental Design

Thread retrieval is a new task, and the number of test collections is limited. In this study, we used the same corpus used by [2]. It has two datasets from two forums—Ubuntu[1] and Travel[2] forums. The statistics of the corpus is as follows. In the Ubuntu dataset, there are 113277 threads, 676777 messages, 25 queries and 4512 judged threads. In the Travel dataset,there are 83072 threads, 590021 messages, 25 queries and 4478 judged threads. The same relevance protocol was followed: a thread with a 1 or a 2 relevance judgment is considered as relevant, while a thread with relevance judgment of 0 is considered as irrelevant. Text was stemmed with the Porter stemmer and no stopword removal was applied. In conducting the experiments, we used the Indri retrieval system[3].

As for evaluation, we used Precision at 10 (P@10), Normalized Discounted Cumulative Gain at 10 (NDCG@10) and Mean Average Precision (MAP). In addition, we used the virtual document model(VD)[4] as our baseline. We used VD because it has been used as a strong baseline in most previous studies[4, 2, 9]. In addition to VD, we used the fused thread search methods as baselines as well.

As for parameter estimation, we have three parameters to estimate: the smoothing parameters μ for the virtual document and the message language models, the size of the initial ranked list of messages R_Q and the number of the top k ranked messages. An exhaustive grid search was applied to maximize MAP using 5-fold cross validation.

Note that we estimated the thread search methods' parameters because our interest is assessing the performance of the meta search techniques on combining the various thread search methods. By optimizing these parameters, we can

[1] `ubuntuforums.org`

[2] `http://www.tripadvisor.com/ShowForum-g28953-i4-NewYork.html`

[3] `http://www.lemurproject.org/indri.php`

estimate the upper limit of the meta search techniques performance. Therefore, a more informed decision about the potential of meta search approach can be made.

5 Experimental Result

Table 1 presents the performance of the meta search techniques on the Ubuntu and the Travel datasets. The letters I, R, and M denote thread search methods based on the initial, the reply or the free type messages respectively. A + symbol between these letters indicates a fusion scenario between these methods. In addition, the acronym VD denotes searching threads using their virtual document representation [4]. From data presented on Table 1, several observations are found.

The first observation is related to the performance of the baselines specially the I, R and M methods. The M method has the best performance on all measures. In addition, the second performing methods are the VD and the R methods; where, R beat VD on P@10 and NDCG@10, but it lost to VD on MAP. Lastly, the I method has the worst performance on all measures.

The second observation is related to the performance of the meta search techniques. Generally, the performance can be divided into three levels: inferior, competitive and superior. The division is based on the ability of these techniques to outperform the best performing baseline method(M). The inferior techniques are CombMIN, CombMED and CombANZ. The only competitive technique is CombMAX. The superior techniques are CombSUM and CombMNZ.

The third observation is related to the performance of the combinations of the thread search methods. In fusing using CombSUM, the combination between the M and the I methods is the best performing fusion scenario with improvements over the best baseline method(M).

6 Discussion

Generally, the application of the meta search techniques to thread retrieval is a successful approach for two reasons. First, it is known that a fusion scenario is successful if its fusion result outperformed its best components [12]. From Table 1, we could see many fusion scenarios were superior to their fused components. The second reason is that some fusion scenarios are able to beat all baseline methods. Particularly, the I+M fusion scenario using CombSUM beat M and VD on all measures. In addition, I+M has better performances on P@10 and NDCG@10. That means not only more relevant threads are found on the top 10 retrieved threads, but also these top 10 threads contained more highly relevant threads positioned at higher ranks.

Nevertheless, a close look at the result described above gives several insights about used thread search methods. Firstly, the good performance of the R and the M methods supports the argument that ranking threads using some of its contents is better than ranking them using all contents. In addition, it supports

Table 1. Retrieval performance of the meta search techniques. The best combinations of the thread search methods for each meta search technique are bolded.

Technique	Combination	Ubuntu Dataset			Travel Dataset		
		MAP	P@10	NDCG@10	MAP	P@10	NDCG@10
Baselines	VD	0.3774	0.4800	0.3549	0.3437	0.4200	0.3284
	I	0.1974	0.3360	0.2630	0.1914	0.3560	0.2658
	R	0.2746	0.4600	0.3676	0.3167	0.5440	0.4205
	M	**0.3665**	**0.5080**	**0.4013**	**0.4046**	**0.6400**	**0.5043**
CombMIN	I+R	0.0573	0.1640	0.1070	0.0825	0.3080	0.2251
	I+M	0.0753	0.1800	0.1179	0.0956	0.3320	0.2468
	R+M	**0.2778**	**0.4680**	**0.3873**	**0.3031**	**0.5320**	**0.3963**
	I+R+M	0.0402	0.1240	0.0900	0.0776	0.2720	0.2000
CombMAX	I+R	0.2746	0.4600	0.3676	0.3168	0.5440	0.4205
	I+M	**0.3665**	**0.5080**	**0.4013**	**0.4046**	**0.6400**	**0.5043**
	R+M	0.3045	0.4640	0.3644	0.4039	0.6320	0.4992
	I+R+M	0.3045	0.4640	0.3644	0.4039	0.6320	0.4992
CombMED	I+R	0.1061	0.2200	0.1759	0.0998	0.3240	0.2341
	I+M	0.1502	0.2600	0.1949	0.1221	0.3520	0.2596
	R+M	**0.2974**	**0.4840**	**0.3924**	**0.3658**	**0.5760**	**0.4472**
	I+R+M	0.2542	0.4280	0.3510	0.2885	0.5320	0.4090
CombSUM	I+R	0.3035	0.4960	0.4008	0.3512	0.6000	0.4667
	I+M	**0.3680**	**0.5400**	**0.4183**	**0.4040**	**0.6480**	**0.5102**
	R+M	0.3165	0.4840	0.3925	0.3687	0.6080	0.4785
	I+R+M	0.3343	0.5120	0.4165	0.3850	0.6280	0.4951
CombMNZ	I+R	0.2979	0.5200	0.4098	0.3355	0.6120	0.4709
	I+M	**0.3600**	**0.5400**	0.4179	**0.3853**	**0.6560**	**0.5103**
	R+M	0.3164	0.4840	0.3925	0.3687	0.6080	0.4785
	I+R+M	0.3316	0.5360	**0.4232**	0.3653	0.6360	0.5003
CombANZ	I+R	0.1061	0.2200	0.1759	0.0998	0.3240	0.2341
	I+M	0.1502	0.2600	0.1949	0.1221	0.3520	0.2596
	R+M	**0.2974**	**0.4840**	**0.3924**	**0.3658**	**0.5760**	**0.4472**
	I+R+M	0.1521	0.2880	0.2273	0.1505	0.3640	0.2698

Table 2. Performance of reply message based thread search methods

Method	MAP	P@10	NDCG@10
	Ubuntu dataset		
R(VD)	0.1766	0.3520	0.2831
R(PCS)	0.2746	0.4600	0.3676
	Travel dataset		
R(VD)	0.2490	0.4160	0.3000
R(PCS)	0.3262	0.5680	0.4433

using the message index to represent threads. However, a counter argument might be that the performance of R might not due to using the message index but due to using only the replies. To test the validity of this argument, a comparison between R using the message index and R using the concatenations of all reply messages [2] was carried out. The result shown in Table 2 rejects this argument. On both datasets,in using the reply messages, the Pseudo Cluster Selection methods are superior to the concatenation of these replies and building a separate index for them. Lastly, the inferiority of the I method to the R, the M and the VD methods implies that relying on a single message to capture a thread relevance is not sufficient.

The low performance of CombMIN, CombMED and CombANZ is due to their sensitivity to the low performing fused components. That explains why the R+M fusion scenario is better than the I+R or the I+M scenarios on these methods. CombMIN will favor the worst among the fused components, while CombMED will promote the middle value — in this case, it is R. On the I+R+M and I+R scenarios, adding up the low scores of the I and the R methods will cancel the high score of M. In contrast, on the R+M scenario, the fusion is between the best and the second best components, hence it leads to a better ranking as the effects of the low scores on ranking slightly demolish.

The performance of CombMAX indicates that M is a strong evidence, and it is dominant over the other methods. In addition, a link between the performance of CombMAX and that of CombSUM and CombMNZ can be found. From CombMAX, we learned that M is a strong evidence; therefore, adding a supportive evidence such as the initial message will likely improve retrieval. Nevertheless, an interesting question that rises is: when fusing using CombSUM, why I+M is better than R+M? Intuitively, one will expect R+M to work better because it involves R, which is better than I. This counter intuitive result is explained by that M and R might use the same content to rank threads. In the M method, suppose a thread was scored with three messages, then at least two of these messages are replies— if not all of them. Therefore, in the R method, these reply

Table 3. Overlap percentage between the used thread search methods

Dataset	Methods	All documents	Relevant	Non relevant
Ubuntu	I,R	0.19	0.70	0.17
	I,M	0.40	0.82	0.38
	R,M	0.54	0.87	0.53
Travel	I,R	0.15	0.60	0.13
	I,M	0.29	0.73	0.27
	R,M	0.67	0.87	0.66

messages might be selected again leading to retrieving the same threads that were retrieved by M. To confirm this premise, we calculated the overlap percentage between the results of these methods. As shown in Table 3, the similarity ratio between I and M is less than it is between R and M.

7 Conclusion and Future Works

In this study, we addressed the problem of searching online forum threads using multiple representations. Current methods that leverage multiple thread representations on searching threads require relevance judgments. However, relevance judgments are hard to obtain on online forums. As a result, we proposed to use meta search techniques to search threads using multiple representations. Many meta search techniques do not require relevance judgments and perform well on many search tasks [6, 11].

Experimenting with three thread representations and six meta search techniques, two findings were found. First, many meta search based fusion scenarios are able to beat their fused components on several measures. Second, using the CombSUM or the CombMNZ to fuse the initial message based method and the free type message based method outperformed all baseline methods on several measures.

Our future work is to fuse more thread search methods such as searching threads using the concatenations of the thread message texts or using the title content. In addition, we will investigate the use of the rank based meta search techniques such as the BordaFuse [1] and the Condorcet [5] methods. Lastly, we will compare the performance of the meta search based fusion to the supervised models proposed by [2, 9].

Acknowledgements. This work is supported by Ministry of Higher Education (MOHE) and Research Management Centre (RMC) at the Universiti Teknologi Malaysia (UTM) under Research University Grant Category (VOT J130000.7 826. 4F011).

References

[1] Aslam, J.A., Montague, M.: Models for metasearch. In: Proceedings of the 24th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR 2001, pp. 276–284. ACM, New York (2001)

[2] Bhatia, S., Mitra, P.: Adopting inference networks for online thread retrieval. In: Proceedings of the Twenty-Fourth AAAI Conference on Artificial Intelligence, Atlanta, Georgia, USA, July 11–15, pp. 1300–1305 (2010)

[3] Elsas, J.L.: Ancestry.com online forum test collection. Technical Report CMU-LTI-017, Language Technologies Institute, School of Computer Science, Carnegie Mellon University (2011)

[4] Elsas, J.L., Carbonell, J.G.: It pays to be picky: an evaluation of thread retrieval in online forums. In: Proceedings of the 32nd International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR 2009, pp. 714–715. ACM, New York (2009)

[5] Montague, M., Aslam, J.A.: Condorcet fusion for improved retrieval. In: Proceedings of the Eleventh International Conference on Information and Knowledge Management, CIKM 2002, pp. 538–548. ACM, New York (2002)

[6] Ogilvie, P., Callan, J.: Combining document representations for known-item search. In: Proceedings of the 26th Annual International ACM SIGIR Conference on Research and Development in Informaion Retrieval, SIGIR 2003, pp. 143–150. ACM, New York (2003)

[7] Ponte, J.M., Croft, W.B.: A language modeling approach to information retrieval. In: Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR 1998, pp. 275–281. ACM, New York (1998)

[8] Seo, J., Croft, W.B.: Blog site search using resource selection. In: Proceedings of the 17th ACM Conference on Information and Knowledge Management, CIKM 2008, pp. 1053–1062. ACM, New York (2008)

[9] Seo, J., Bruce Croft, W., Smith, D.: Online community search using conversational structures. Information Retrieval 14, 547–571 (2011)

[10] Shaw, J.A., Fox, E.A., Shaw, J.A., Fox, E.A.: Combination of multiple searches. In: The Second Text REtrieval Conference (TREC-2), pp. 243–252 (1994)

[11] Wu, M., Hawking, D., Turpin, A., Scholer, F.: Using anchor text for homepage and topic distillation search tasks. Journal of the American Society for Information Science and Technology 63(6), 1235–1255 (2012)

[12] Wu, S.: Data Fusion in Information Retrieval. Springer, Heidelberg (2012)

[13] Zhai, C., Lafferty, J.: A study of smoothing methods for language models applied to information retrieval. ACM Trans. Inf. Syst. 22(2), 179–214 (2004)

Ensemble of Diversely Trained Support Vector Machines for Protein Fold Recognition

Abdollah Dehzangi[1,2] and Abdul Sattar[1,2]

[1] Institute for Integrated and Intelligent Systems (IIIS),
Griffith University, Brisbane, Australia
[2] National ICT Australia (NICTA), Brisbane, Australia
{a.dehzangi,a.sattar}@griffith.edu.au

Abstract. Protein Fold Recognition (PFR) is defined as assigning a given protein to a fold based on its major secondary structure. PFR is considered as an important step toward protein structure prediction and drug design. However, it still remains as an unsolved problem for biological science and bioinformatics. In this study, we explore the impact of two novel feature extraction methods namely overlapped segmented distribution and overlapped segmented autocorrelation to provide more local discriminatory information for the PFR compared to previously proposed methods found in the literature. We study the impact of our proposed feature extraction methods using 15 promising physicochemical attributes of the amino acids. Afterwards, by proposing an ensemble Support Vector Machines (SVM) which are diversely trained using features extracted from different physicochemical-based attributes, we enhance the protein fold prediction accuracy for up to 5% better than similar studies found in the literature.

Keywords: Overlapped segmented distribution, Overlapped segmented autocorrelation, Physicochemical-based features, Ensemble of different classifiers, Support vector machine.

1 Introduction

Protein Fold Recognition (PFR) is considered as an important step towards protein structure prediction problem and drug design. It provides critical information of the classification of proteins based on their general major secondary structure. In pattern recognition perspective, PFR is defined as solving a multi-class classification task which its prediction performance depends on features and classification techniques being used. During the past two decades, a wide range of classification techniques such as: *Artificial Neural Network (ANN)* [1,2], Meta Classifiers [3, 4], and *Support Vector Machines (SVM)* [5] have been used to solve this problem. Among the employed classifiers, ensemble of different classifiers attained the best results for PFR [6, 7]. Similar to exploring the classification techniques, a wide range of features have been proposed and used to solve this problem. Features being using for PFR can be

A. Selamat et al. (Eds.): ACIIDS 2013, Part I, LNAI 7802, pp. 335–344, 2013.

generally categorized into three groups namely, physicochemical-based features (extracted from the physicochemical properties of the amino acids (e.g. hydrophobicity) [8–10]), sequential-based features (extracted from the alphabetic sequence of proteins (e.g. occurrence of amino acids [11])), and evolutionary-based features (extracted from the *Position Specific Scoring Matrix (PSSM)* [12]). Among these groups, physicochemical-based features are the only group that maintain their discriminatory information when the sequential similarity rate is low. Therefore, they attract tremendous attention for PFR [4, 7, 13].

To explore the impact of different physicochemical-based attributes for PFR, Gromiha and his co-workers [8] studied 49 different properties of the amino acids. they extracted 49 features based on the global density of these attributes (49-D feature vector) and used for the PFR. Despite a wide range of attributes explored in this study, they failed to properly explore the local discriminatory information of these attributes (due to use of a feature that solely describe the global density of a given attribute). To explore more local information for PFR, later studies shifted the focus to more sophisticated feature extraction methods [9,14]. However, they merely relied on a few popular physicochemical-based attributes for feature extraction (e.g. hydrophobicity, polarity, flexibility). Furthermore, in all these studies, the whole protein sequence used as a building block of extracting local discriminatory information. Therefore, they failed to provide adequate local discriminatory information for large proteins.

In this study, we aim at enhancing the protein fold prediction accuracy by addressing the limitations highlighted earlier in the following two steps. First, to provide more local information compared to previously explained methods, we propose two feature extraction methods namely, overlapped segmented distribution and overlapped segmented autocorrelation. We explore the impact of our proposed feature extraction method for 15 most promising physicochemical-based attributes using five classification techniques that attained good results for PFR (Adaboost.M1, SVM, Random Forest, Naive Bayes, and *Ensemble of Different Classifiers (EDC)* proposed in [2]). Then, by proposing an ensemble of diversely trained SVM classifiers using features extracted from different physicochemical-based attributes, we enhance protein fold prediction accuracy up to 5% better than similar studies found in the literature.

2 Data Sets and Features

2.1 Data Sets

To investigate the performance and generality of our proposed methods, two datasets namely, EDD (extended version of the DD dataset introduced by [9]) and TG (introduced by [11]) are used in this study. We extract the EDD data set from the latest version of *Structural Classification of Proteins (SCOP)* (1.75) in the similar manner used in [5] to replace the old DD dataset which no longer used due to its inconsistency with the SCOP 1.75. This dataset consist of 3418 proteins with less than 40% sequential similarity belonging to 27 folds as it was used in DD. The EDD dataset is mainly used to investigate the performance of our proposed method compared to similar studies found in the literature. We also used the TG benchmark extracted by [11] from the SCOP 1.73 consists of 1612 proteins with less than 25% sequential similarities belonging to 30 most populated folds in SCOP. The TG benchmark is mainly used to investigate the performance of our proposed approaches when the sequential similarity rate is low. To simulate the DD dataset condition and to be able to directly compare our results with previous studies, we divided each of our employed datasets into train and test sets (in the manner that 3/5 of the data is used in training set and 2/5 of the data is used in the testing set [9]).

2.2 Physicochemical-Based Attributes

In this study, we explored the impact of our proposed approaches for 15 most promising physicochemical-based attributes. These attributes were selected by the authors from a wide range of physicochemical-based attributes explored experimentally. We studied the performances of features extracted from 115 different physicochemical-based attributes (extracted mainly from the APD database [15] and the [8] study) and selected the following 15 attributes: (1) structure derived hydrophobicity value, (2) polarity, (3) average long range contact energy, (4) average medium range contact energy, (5) mean RMS fluctuational displacement, (6) total non-bounded contact energy, (7) amino acids partition energy, (8) normalized frequency of alpha-helix, (9) normalized frequency of turns, (10) hydrophobicity scale derived from 3D data, (11) HPLC parameters to predict hydrophobicity and antigenicity, (12) average gain ratio of surrounding hydrophobicity, (13) mean fractional area loss, (14) flexibility, and (15) bulkiness. Note that to the best of our knowledge, most of the selected attributes (attributes number 3, 4, 5, 6, 7, 10, 11, 12, 13, and 14) have not been adequately (or not at all) explored for the PFR. However, the conducted comprehensive experimental study showed they are able to outperformed many popular attributes that have been widely used for PFR [1, 7, 9, 10].

3 Proposed Feature Extraction Methods

In this study, we propose two novel feature extraction methods namely overlapped segmented distribution and overlapped segmented autocorrelation. The proposed feature extraction methods are aimed at providing more local discriminatory information than previously used approaches found in the literature. These approaches are discussed in the following subsections.

3.1 Overlapped Segmented Distribution

As it was highlighted earlier, previously, global density was used as descriptor of a given physicochemical-based attributes in [8]. To calculate this feature, the amino acids in a given protein sequence (A1,A2, ..., AL where L is the length of the protein) is first replaced with their numerical values assigned to them based on a given attribute (R1,R2, ..., RL). Then it is calculated as follow:

$$T_{glob\_dens} = \frac{\sum_{i=1}^{L} R_i}{L}. \tag{1}$$

However, it could not properly explore the discriminatory information, embedded locally in a given physicochemical-based attribute. To address this limitation, we propose overlapped segmented-based distribution feature set. This feature set is calculated as follows. Beginning from the left side of a given protein, we sum the attribute values of the amino acids until reaching to *K%* of the total sum (which is equal to $T_{sum} = T_{glob\text{-}dens} \times L$) as follows:

$$C_k^l \leq (T_{sum} \times K)/100. \tag{2}$$

Then, the number of summed amino acids divided by the length of the protein is returned as the distribution of the first K% of the global density. We repeat this process for 2K, 3K, ... , nK, calculate the $C_{2K}^l, C_{3K}^l, \dots, C_{nK}^l$ and return the given distribution-based features accordingly (where nK = 75). Same process is conducted from the right side. We start from right side and for K, 2K, ... ,nK, we calculate the $C_{2K}^r, C_{3K}^r, \dots, C_{nK}^r$ and return the corresponding distributionbased features (Figure 1). Therefore, *75/K* features (*n=75/K*) from each side are extracted in this feature set (totally *150/K* = *75/K* × *2*). We also added the global density feature to this features as a global descriptor (*150/K* + *1* features in total). In this study, the distribution factor K=5 is adopted due to its better performance attained experimentally compared to other distribution factors such as *K* = *10* and *K* = *25* [9,10]. We also adopt 75 as the overlapping factor which showed better performance with respect to the number of features generated compared to other factors. Hence 31 features extracted in this feature set (*150/5* + *1* = *31*).

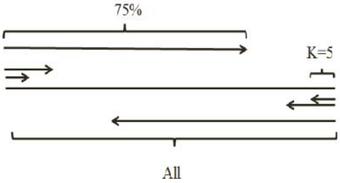


Fig. 1. Segmented distribution-based feature extraction method

In this method, we calculated the distribution factor from both side of a protein sequence to bring the emphasize to the sides of proteins. To highlight the impact of the middle part of a protein sequence as well, we adopt overlapping style. In this manner, the impact of a give attributes with respect to each side of the protein sequence are represented.

3.2 Overlapped Segmented Autocorrelation

The overlapped segmented-based distribution which was introduced earlier, is mainly based on the density and distribution properties. In this section, we propose overlapped segmented-based autocorrelation which is based on the autocorrelation property. Autocorrelation of the amino acids have been widely used as an effective feature which reveals important information of how amino acids are ordered in the protein sequence [14]. However, previous approaches (even the most sophisticated ones (e.g. pseudo amino acid composition [14])) failed to properly explore the potential of this method [7].

Therefore, in this study, we propose the concept of segmented autocorrelation aiming to address this limitation. These feature set is extracted in the following manner. We first segments the protein sequence as it used in the overlapping segmentation-based distribution method (where segmentation factor *K* = *10* and overlapping factor 70 is adopted due to better performance attained experimentally). Then we calculate the autocorrelation with distance factor *D_F* = *10* (as it was shown the most effective value for this parameter in [16]) cumulatively. In the similar manner to overlapped segmented-based distribution method, we calculate these features starting from each side (left and right). We therefore calculate *T_F* =7 × 10 features from each side (totally 140 = *T_F* × 2 from both sides). We also added the autocorrelation calculated using the whole protein sequence as the global descriptor of this method (totally 150 = *140* + *10* features). The autocorrelation in each segment is equal to:

$$A\_Ci, a = \frac{1}{(L \times (a/100) - i)} \sum_{j=m}^{n} S_j \times S_{j+i}, (i = 1, ..., 10 \,\&\, a = 10, .., 70), \tag{3}$$

where a is the segmentation factor, m and n are respectively the begin and the end of a segment, and Sj is the attribute value for each amino acid.

4 Classification Techniques

In this study, we use four different classifiers namely, AdaBoost.M1, Random Forest, Naive Bayes, and SVM to evaluate the performance of the explored physicochemical-based attributes with respect to the proposed feature extraction methods. These classifiers are selected based on their performances attained in the previous studies for the PFR [2–5,16]. These classifiers are briefly introduced as follows.

Naive Bayes: As the most popular Bayesian-based classifier is based on the naive assumption of independency of the studied features for a given task. Despite its simplicity, it attained promising results for this task [2]. Naive Bayes is also able to provide important information of the correlation between the features being used (better performance of Naive Bayes is the prove for low level of correlation between features while poor performance will support the high rate of correlation between them) [2].

AdaBoost.M1: Is introduced by [17] and is considered as the best-of-the-shelf meta classifier. Adaboost.M1 also attained promising results for the PFR [2]. It uses a classifier called base learner sequentially in K iteration and adopt the weight of the misclassified samples in each iteration to improve the performance. It builds its final structure by combining the output of each step for a given sample using majority voting as its final decision. In this study, Adaboost.M1 implemented in WEKA is used [18]. The C4.5 decision tree is used as its base learner and the number of its iterations is set to 100 ($K=100$) (as it is shown as the best parameter for this algorithm for the PFR [19].

Random Forest: Is introduced by [20] based on the concept of bagging. It randomly select K subset of features and train K different classifiers (base learners) independency and then, combine their results using majority voting as its final decision. Despite its simplicity, it was shown as an effective classifier for different tasks as well as the PFR [3]. In this study, for the Random Forest (implemented in WEKA) $K = 100$ and random tree based on the gain ratio is used as its base learner.

Support Vector Machine: Is considered as the most promising classification technique which outperformed other individual classifiers for the PFR [5,16]. SVM aims as finding the *Maximal Marginal Hyperplane (MMH)* to minimizing the classification error. To find the appropriate support vector, it transforms the input data to different dimension based on the concept of kernel function. In this study, We employ SVM using *Sequential Minimal Optimization (SMO)* as a kind of polynomial kernel (implemented in WEKA) which its kernel degree is set to one ($p = 1$).

5 Results and Discussion

In the fist step, we construct a feature vector based on each attribute explored in this study with respect to the proposed feature extraction methods. Therefore, for a given attribute a feature vector consists of two feature groups namely overlapped segmented distribution feature group (31 features) and overlapped segmented autocorrelation (150 features) is constructed. We also added the composition of the amino acids feature group (20 features) as well as the length of the protein sequence feature (1 feature) to these feature vectors as important sources of sequential-based features [1, 19]. Composition of the amino acid feature group consists of the percentage of occurrence of amino acids divided by the length of protein sequence which used in [9] and later works as an effective feature group that provide important sequential-based information. The length of the amino acids also showed as an effective features for the PFR [2,4]. Therefore, based on each attribute being explored in this study, a feature vector consist of 202 features (31 + 150 + 20 + 1) constructed to explore their local and global discriminatory information in detail. For the rest of this study, each feature group will be referred as *comb numb* where *numb* is the number assigned to the corresponding physicochemical-based attribute in section 2.2.

We then apply the employed classifiers in this study (Naive Bayes, AdaBoost.M1, SVM, and Naive Bayes) to the constructed feather vectors and report the results in Table 1. We also study the performance of the *Ensemble of Different Classifier (EDC)* which we introduced in our previous work [2] to the extracted feature vectors to compare its results with the results achieved in previous studies. We first reproduce the results achieved in [10] using EDC (219 features) and achieve up to 48.8% and 41.1% prediction accuracies respectively for the EDD and TG datasets. We also reproduce the results using EDC for the features introduced in [9] (126 features) and 69D feature vector (49D in addition to 20 features for the composition feature group) and achieve to 47.6% and 36.6% for EDD and 40.7% and 33.0% for TG respectively.

Table 1. Results achieved (in percentage %) by using AdaBoost.M1 (Ada), Random Forest (RF), Naive Bayes (NA), SVM, and EDC for 15 feature vectors extracted from the explored attributes in this study (for both EDD and TG datasets)

Datasets	EDD					TG				
Comb_Numb	Ada	Na	RF	SVM	EDC	Ada	Na	RF	SVM	EDC
Comb_1	45.8	24.9	40.6	50.1	50.3	37.4	22.2	33.1	37.7	39.3
Comb_2	45.4	28.8	38.5	49.7	51.1	35.7	24.0	31.7	40.9	39.3
Comb_3	45.8	26.3	41.2	49.6	49.5	38.6	22.4	34.9	41.3	40.7
Comb_4	43.9	25.6	39.3	47.9	46.7	35.3	24.5	34.4	37.9	37.9
Comb_5	42.9	24.4	40.3	50.5	50.0	35.8	17.5	31.3	39.8	39.1
Comb_6	45.1	19.5	39.9	51.3	50.5	39.1	17.5	34.7	42.4	**43.9**
Comb_7	42.6	23.0	38.9	49.2	49.0	35.5	17.5	30.6	39.4	39.0
Comb_8	45.8	29.0	40.6	48.1	48.6	36.0	23.0	36.3	39.1	40.7
Comb_9	43.0	23.3	39.8	48.1	47.1	35.8	17.8	32.8	40.2	39.3
Comb_10	45.6	26.9	38.3	52.4	52.0	36.3	22.1	33.0	41.0	42.7
Comb_11	43.4	25.0	39.4	50.6	50.1	33.9	20.7	32.0	42.1	40.4
Comb_12	43.3	24.0	39.2	49.7	49.7	38.5	20.8	35.2	40.4	40.5
Comb_13	43.9	23.2	38.6	52.4	**52.8**	37.7	18.9	34.2	40.5	41.6
Comb_14	43.6	24.2	38.3	50.7	50.2	36.8	20.0	31.7	40.4	42.3
Comb_15	42.9	19.1	38.4	48.1	48.1	35.3	17.7	34.2	40.5	39.1

As it is shown in Table 1. By using EDC to the Comb 13 feature vector (mean fractional area loss attribute which to the best of our knowledge, have not been used for the PFR), we achieve to 52.8% prediction accuracy, up to 4% better than previously reported results in similar studies for the EDD dataset. We also achieve up to 43.9% prediction accuracy using EDC to the Comb 6 feature vector (total non-bounded contact energy attribute which to the best of our knowledge, have not been adequately explored for the PFR), over 2.8% better than previously reported results in similar studies for the TG dataset. Our results emphasize on the effectiveness of our proposed feature extraction methods as well as the importance of physicochemical-based attributes explored in this study against the popular feature extraction methods and attributes have been used for PFR. As it is shown in Table 1, the enhancement for the EDD and TG datasets are achieved using different feature vectors. Hence, we also propose a fusion of diversely trained ensemble of SVM to explore the potentials of the explored attributes in conjunction with each other as well as defining a system that generally perform well for both EDD and TG benchmarks.

To achieve this goal, we propose an ensemble of four diversely trained SVM classifiers. SVM is selected due its it better performance compared to the other employed classifiers in this study as well as its promising performance for the PFR. We trained these four SVM classifiers with four feature vectors extracted from four physicochemical-based attributes that attained the best results mainly for TG dataset among the explored attributes in this study (which raises the idea of exploring more optimal approach for this task in future studies.). We fuse these classifiers and use EDC (trained on Comb 10 which attained persistent results for both EDD and TG datasets) as a tie breaker in our proposed system. The voting system works in the following manner. For the case that majority of SVM classifiers, classify a given sample to a same fold, this fold will also be chosen as the decision and consequently output fold without consideration of the output of the EDC. In case that not a single fold would have the majority votes, the output of the EDC will be chosen as the output of the system (which occur when two of the SVM classifiers vote for a same fold and other two SVM classifiers for another fold or all four SVM classifiers vote for different folds). The architecture of our proposed *Ensemble of Diversely Trained SVM Classifiers (EDTSVM)* is shown in Figure 2.

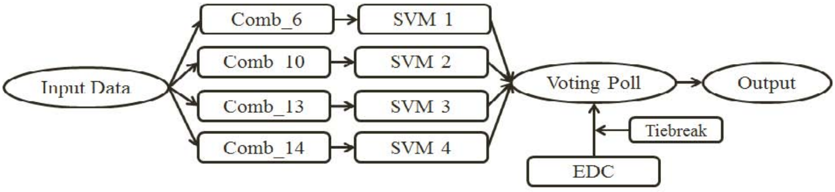


Fig. 2. The overall architecture of the EDTSVM

Table 2. The best results (in percentage %) achieved in this study compared to the best results found in the literature for the EDD and TG benchmarks respectively

Study	Attributes (Number of features)	Method	EDD (Results)	TG (Results)
[9]	Features proposed in [9] (126)	SVM	46.3	38.5
[19]	Features proposed in [9] (126)	Ada	44.7	36.4
[3]	Features proposed in [9] (126)	RF	42.9	37.1
[2]	Features proposed in [9] (126)	EDC	47.6	40.7
[2]	Features proposed in [10] (219)	SVM	47.3	40.1
[19]	Features proposed in [10] (219)	Ada	45.3	37.2
[3]	Features proposed in [10] (219)	RF	43.9	38.1
[2]	Features proposed in [10] (219)	EDC	48.8	41.1
[8]	69D (49+20)	SVM	36.6	33.0
This study	Comb 6 (202)	SVM	51.3	42.4
This study	Comb 10 (202)	SVM	52.4	41.0
This study	Comb 11 (202)	SVM	50.6	42.1
This study	Comb 6 (202)	EDC	50.6	43.9
This study	Comb 10 (202)	EDC	52.0	42.7
This study	Comb 13 (202)	EDC	52.8	41.6
This study	Fused (202 for each classifier)	EDTSVM	53.8	43.5

By applying the EDTSVM to the EDD and TG datasets, we achieve up to 53.8% and 43.5% prediction accuracies, up to 5% and 2.4% better than the best results reported in similar studies found in the literature. We also achieved up to 17.2% and 10.5% better prediction accuracy than using 69D for the EDD and TG datasets, respectively. These highlights the impact of the proposed feature extraction methods in this study to reveal significant discriminatory information based on an individual attribute rather than using a naive feature extraction method for a wide range of physicochemical-based attributes. The comparison of the results achieved in this study, compared to the similar studies found in the literature for the TG and EDD benchmarks is shown in Table 2.

5.1 Conclusion and Future Works

In this study, we explored the impact of 15 physicochemical-based attributes using two novel feature extraction methods namely, overlapping segmented-based distribution and overlapping segmented-based autocorrelation. We then, constructed a feature vector consisting of combination of features extracted using our feature extraction methods as well as composition of the amino acids feature group and the length of protein sequence feature. Then by using several classifiers that attained good results for the PFR such as, Random Forest, AdaBoost.M1, Naive Bayes, SVM, and EDC (proposed in [2]), we studied the effectiveness of our proposed approaches. Achieved results showed the impact of our proposed feature extraction methods with respect to the attributes being used compared to the similar studies found in the literature. Finally, By proposing an ensemble of diversely trained SVM classifiers (EDTSVM) applied to the feature vectors extracted with respect to the physicochemical-based attributes that have not been adequately explored for PFR, we achieved up to 5% prediction accuracy compared to the similar studies found in the literature.

For our future works, we aim to explore the impact of evolutionary-based information in conjunction with physicochemical-based information for PFR. We also aim to explore the impact of weighted ensemble of different classifiers based on the classifier as well as features being used.

References

1. Ghanty, P., Pal, N.R.: Prediction of protein folds: Extraction of new features, dimensionality reduction, and fusion of heterogeneous classifiers. IEEE Transactions on NanoBioscience 8(1), 100–110 (2009)
2. Dehzangi, A., Phon Amnuaisuk, S., Ng, K.H., Mohandesi, E.: Protein Fold Prediction Problem Using Ensemble of Classifiers. In: Leung, C.S., Lee, M., Chan, J.H. (eds.) ICONIP 2009, Part II. LNCS, vol. 5864, pp. 503–511. Springer, Heidelberg (2009)
3. Dehzangi, A., Phon-Amnuaisuk, S., Dehzangi, O.: Using random forest for protein fold prediction problem: An empirical study. Journal of Information Science and Engineering 26(6), 1941–1956 (2010)

4. Chen, K., Kurgan, L.A.: Pfres: protein fold classification by using evolutionary information and predicted secondary structure. Bioinformatics 23(21), 2843–2850 (2007)
5. Yang, J.Y., Chen, X.: Improving taxonomy-based protein fold recognition by using global and local features. Proteins: Structure, Function, and Bioinformatics 79(7), 2053–2064 (2011)
6. Dehzangi, A., Karamizadeh, S.: Solving protein fold prediction problem using fusion of heterogeneous classifiers. INFORMATION, An International Interdisci¬plinary Journal 14(11), 3611–3622 (2011)
7. Yang, T., Kecman, V., Cao, L., Zhang, C., Huang, J.Z.: Margin-based ensemble classifier for protein fold recognition. Expert Systems with Applications 38, 12348–12355 (2011)
8. Gromiha, M.M., Oobatake, M., Sarai, A.: Important amino acid properties for enhanced thermostability from mesophilic to thermophilic proteins. Biophysical Chemistry 82, 51–67 (1999)
9. Ding, C., Dubchak, I.: Multi-class protein fold recognition using support vector machines and neural networks. Bioinformatics 17, 349–358 (2001)
10. Dehzangi, A., Phon-Amnuaisuk, S.: Fold prediction problem: The application of new physical and physicochemical-based features. Protein and Peptide Letters 18(2), 174–185 (2011)
11. Taguchi, Y.H., Gromiha, M.M.: Application of amino acid occurrence for discrimi¬nating different folding types of globular proteins. BMC Bioinformatics 8(1), 404 (2007)
12. Kurgan, L.A., Cios, K.J., Chen, K.: Scpred: Accurate prediction of protein struc¬tural class for sequences of twilight-zone similarity with predicting sequences. BMC Bioinformatics 9, 226 (2008)
13. Kavousi, K., Moshiri, B., Sadeghi, M., Araabi, B.N., Moosavi-Movahedi, A.A.: A protein fold classifier formed by fusing different modes of pseudo amino acid composition via pssm. Computational Biology and Chemistry 35(1), 1–9 (2011)
14. Shen, H.B., Chou, K.C.: Ensemble classifier for protein fold pattern recognition. Bioinformatics 22, 1717–1722 (2006)
15. Mathura, V.S., Kolippakkam, D.: Apdbase: Amino acid physico-chemical properties database. Bioinformation 12(1), 2–4 (2005)
16. Dong, Q., Zhou, S., Guan, G.: A new taxonomy-based protein fold recognition approach based on autocross-covariance transformation. Bioinformatics 25(20), 2655–2662 (2009)
17. Freund, Y., Schapire, R.E.: Experiments with a new boosting algorithm. In: International Conference on Machine Learning, pp. 148–156 (1996)
18. Witten, I., Frank, E.: Data Mining: Practical Machine Learning Tools and Tech¬niques, 2nd edn. Morgan Kaufmann, San Francisco (2005)
19. Krishnaraj, Y., Reddy, C.K.: Boosting methods for protein fold recognition: An empirical comparison. In: Proceedings of the 2008 IEEE International Conference on Bioinformatics and Biomedicine, pp. 393–396 (2008)
20. Breiman, L.: Bagging predictors. Machine Learning 24, 123–140 (1996)

Protein Fold Recognition Using Segmentation-Based Feature Extraction Model

Abdollah Dehzangi[1,2] and Abdul Sattar[1,2]

[1] Institute for Integrated and Intelligent Systems (IIIS), Griffith University, Brisbane, Australia
[2] National ICT Australia (NICTA), Brisbane, Australia
{a.dehzangi,a.sattar}@griffith.edu.au

Abstract. Protein Fold recognition (PFR) is considered as an important step towards protein structure prediction. It also provides significant information about general functionality of a given protein. Despite all the efforts have been made, PFR still remains unsolved. It is shown that appropriately extracted features from the physicochemical-based attributes of the amino acids plays crucial role to address this problem. In this study, we explore 55 different physicochemical-based attributes using two novel feature extraction methods namely segmented distribution and segmented density. Then, by proposing an ensemble of different classifiers based on the AdaBoost.M1 and Support Vector Machine (SVM) classifiers which are diversely trained on different combinations of features extracted from these attributes, we outperform similar studies found in the literature for over 2% for the PFR task.

Keywords: Segmented distribution, Segmented Density, Physicochemical-based features, SVM, AdaBoost.M1, Ensemble of different Classifiers.

1 Introduction

Determining how a given protein is categorized to a fold based on its major secondary structure is called *Protein Fold Recognition (PFR).* PFR is considered as an important step toward protein structure prediction. It also can provide significant information about general functionality of proteins. During the past few decades, a wide range of approaches proposed to solve PFR mainly based on classification techniques [1–4] as well as feature extraction methods [5–8]. Among the classification techniques used to tackle this problem, ensemble-based classifiers attained the best results for PFR [2, 9, 10]. They outperformed individual classifier used for this task which have driven the focus to these techniques [9–11].

Beside classification techniques used to approach PFR, feature extraction have also attained tremendous attention. Among the features being used for this task, physicochemical-based features (extracted based on physicochemical attributes of the amino acids (e.g. hydrophobicity)) showed promising results. It was shown that dissimilar to the other features used to tackle this problem (e.g. sequential

A. Selamat et al. (Eds.): ACIIDS 2013, Part I, LNAI 7802, pp. 345–354, 2013.

based features which are extracted from the alphabetic sequence of the amino acids) physicochemical-based features maintain their discriminatory information when the sequential similarity rate is low. To the best of our knowledge, The impact of the widest range of physicochemical-based attributes for PFR was explored by Gromiha and his co-workers [7]. They explored 49 different physicochemical attributes of the amino acids using global density feature. Therefore, they extracted 49 features from these attributes. However, due to use of global density for feature extraction which just generate one global feature, they could not explore the local impact of the explored features properly.

To address this issue, later studies shifted their focus to explore fewer number of physicochemical-based attributes and instead, use more efficient feature extraction methods [5,12]. Recent studies shifted the focus to increase the number of attributes being explored as well as providing adequate local discriminatory information by categorizing amino acids into several subgroups based on the concept of alphabet reduction [8]. However, due to use of alphabet reduction, they discarded important information and could not appropriately enhance PFR. Furthermore, similar to their previous works, they tried to extract local information from the whole protein sequence as a single building block which failed to work properly, specially for large proteins.

In this study, we explore 55 different physicochemical-based attributes for PFR. To the best of our knowledge, most of these attributes have not been adequately explored for this task. We also propose two novel segmented base feature extraction methods which are aimed to provide more local discriminatory information than previously proposed approaches for PFR. We explore the impact of our propose approaches using four popular classification techniques namely, *AdaBoost.M1, SVM (SVM), Random Forest*, and *Naive Bayes* which have attained promising results for this task. In the final step, by proposing an ensemble of different classifiers (based on AdaBoost.M1 and SVM) which are diversely trained with the features extracted from a wide range of physicochemical-based attributes, we enhance the protein fold prediction accuracy for more than 2% better than similar studies found in the literature.

2 Datasets and Physicochemical-Based Attributes

In this study, two popular benchmarks namely EDD (extended version of the DD introduced by [5]), and TG (introduced by [13])are used. To be able to directly compare our results with the similar studies found in the literature, the EDD data set is used. We extract this data set from the latest version of the *Structural Classification of Proteins (SCOP 1.75)* consisting of 3418 proteins with less than 40% sequential similarities belonging to 27 fold used previously in DD (similar to [11]). We also use TG benchmark to be able to explore the impact of our proposed approaches when the sequential similarity rate is low. This dataset consists of 1612 proteins with less than 25% sequential similarities belonging to 30 folds. Similar to DD dataset which consists of two separate training and testing sets, we randomly separate the proteins in these datasets to training (3/5

Table 1. Names and number of the explored attributes in this study

No.	Attributes	No.	Attributes
1	Structure derived hydrophobicity value	29	Absolute entropy
2	Polarizability	30	Entropy of formation
3	Normalized frequency of α-helix	31	Buried and accessible molar fraction ratio
4	Normalized frequency of β-strand	32	Energy of transfer from inside to outside
5	Normalized frequency of βturn	33	Flexibility for one rigid residue
6	Hydrophobicity at ph 7.5 by HPLC	34	Side chain interaction parameter
7	Size	35	Side chain volume
8	Consensus normalized hydrophobicity scale	36	Hydropathy index
9	Hyd. index base on helix in membrane	37	Average surrounding hydrophobicity
10	Molecular weight	38	Average reduced distance for side chain
11	Hydrophobic parameter	39	Side chain orientation angle
12	Van Der Waals volume	40	Ave number of nearest neighbor in chain
13	Polarity (driven from amino acids)	41	Average Volume of surrounding residues
14	Volume	42	Hyd. scale (contact energy in 3D data)
15	Compressibility	43	Partition coefficient
16	Average long range contact energy	44	Average gain in surrounding hydrophobicity
17	Average medium range contact energy	45	Surrounding hydrophobicity in α-helix
18	Long range non bounded energy	46	Surrounding hydrophobicity in β-sheet
19	Mean RMS fluctuational displacement	47	Surrounding hydrophobicity in βturn
20	Refractive index	48	Surrounding hydrophobicity in folded form
21	Solvent accessible reduction	49	Average number of surrounding residues
22	Total non bounded energy	50	Membrane buried helix parameter
23	Unfolding entropy change of hydration	51	Mean fractional area loss (f)
24	Unfolding hydration heat capacity change	52	Flexibility
25	Retention coefficient (PH = 7.0)	53	Hydration potential (PH = 7.0)
26	Amino acids partition energy	54	Bulkiness
27	PKa-COOH	55	Polarity (driven from amino acids in proteins)
28	Hyd. value (driven from free amino acids)	-	-

of total proteins) and testing (2/5 of total proteins) to be able to simulate DD dataset's condition.

We also study 55 different physicochemical-based attributes as listed in Table 1 and explore their effectiveness on PFR. These attributes are taken from the APD database [14], and Gromiha and his co-workers study [7]. Our aim in this part is to explore the potential of each attribute to enhance PFR performance with respect to the feature extraction methods being used.

3 Physicochemical-Based Feature Extraction Approaches

In this study, we propose two novel feature extraction methods namely segmented-based density and segmented-based distribution. Our propose approaches are aimed to capture more local discriminatory information compared to previously proposed approaches [5]. These approaches are discussed in the following subsections.

3.1 Segmented Density

This method is mainly proposed to add more local discriminatory information based on the density of a given attribute. In this approach, we replace the amino acids in the original protein sequence (A_1, A_2, ..., A_L where L is the length of the protein) by the attribute values (R_1, R_2, ..., R_L) assigned to the amino acids (e.g. hydrophobicity). Then we segment the protein sequence and calculate the density for each segment. In this study, $K = 5$ segmentation factor is used due to its better performance compared to use of $K = 10$ and $K = 25$ explored experimentally. Hence, protein sequence divided to 20 segments. The expression for segmented density for each segment can be given as follows:

$$SD_{segmented\_density} = \frac{\sum_{i=1}^{D} R_i}{D}, \tag{1}$$

where D $(= L \times (5/100))$ is the length of each segment. Therefore, 20 segmented density features are extracted based on the given method. We also added the global density to these features to add global information to this feature set(20 + 1 features). The expression for global density is given as follows:

$$D_{glob\_density} = \frac{\sum_{i=1}^{L} R_i}{L}. \tag{2}$$

3.2 Segmented Distribution

as it is shown in previous subsection, in segmented density, the segments has equal length. Therefore, the length of segments vary crucially relying on the length of proteins. In this section, we propose a novel feature extraction method based on the concept of segmented-based distribution. In this method, we first calculate the total sum of attribute values (e.g. hydrophobicity) over a given protein sequence which is equal to $T = \sum_{i=1}^{L} R_i$. Then starting from the left side of the protein sequence, we sum the attributes values of the first $I_k^{(l)}$ amino acids until reaching to $K\%$ of T $(T_{seg} \leq (T \times K)/100)$. Then we return the distribution feature of this segment as $I_k^{(l)}/L$. We repeat this procedure for *2K, 3K, ... ,* until reaching to $N \times K = 50$ and calculate the $I_{2k}^{(l)}, I_{2k}^{(l)}, ..., I_{50}^{(l)}$ and then return the $I_{2k}^{(1)}/L, I_{2k}^{(1)}/L, ..., I_{50}^{(1)}/L$ as the assigned distribution features, respectively. The same procedure is done from the right sight to calculate $I_{2k}^{(r)}, I_{2k}^{(r)}, ..., I_{50}^{(r)}$ and then return the $I_{2k}^{(r)}/L, I_{2k}^{(r)}/L, ..., I_{50}^{(r)}/L$ as the assigned distribution features, respectively. Therefore, totally $N_{feat} = 2 \times (50/K) = 100/K$ features are extracted based on a given K in this feature set. The distribution factor *(*K) is a parameter which is determined here experimentally. For this, three values of K (5, 10, and 25) are investigated. To this set of *100/K* distribution features, we add the global density feature to provide more global information. Therefore, we have a total of $N_{feat} + 1$ features. Thus there will be 21, 11, and 6 features for *K=5, 10* and *25*, respectively.

Our proposed physicochemical-based feature extraction methods have two main contributions. First, they provide more local discriminatory information compared to previously adopted methods [7]. Second, instead of categorizing amino acids based on a given attributes to sub groups (as it was adopted in [5]), they work directly with the attributes values assigned to the amino acids. Therefore, they avoid information loss due to alphabet reduction.

4 Classification Techniques

In this study, four classifiers namely, AdaBoost.M1, Naive Bayes, Random Forest, and SVM that attained promising results for PFR used to evaluate the performance of the explored attributes with respect to our proposed feature extraction methods [4, 5, 9]. These classifiers are briefly described as follows:

Naive Bayes: As a kind of a Baysian-Based learner is considered as one of the simplest classifiers yet attained promising results for different tasks as well as PFR [2]. Naive Bayes is based on the assumption of independency of the employed features from each other to calculate the posterior probability [2].

AdaBoost.M1: Is considered as the best-of-the-shelf meta-classifier introduced by [15]. The main idea of the AdaBoost.M1 is to sequentially (in I iterations) apply a base learner (also called weak learner which refer to a classifier that at least performs better than random guess) on the bootstrap samples of data, adjust the weight of misclassified samples, and enhance the performance in each step. In this study, Adaboost.M1 implemented in WEKA using C4.5 decision tree (number of base learners is set to 100 (*I=100*)) as its base learners is employed [16].

Random Forest: Is also considered as a kind of meta-learner which recently attracted tremendous attention specifically for PFR [4]. Random Forest is based on bagging approaches [17]. It applies a base learner independently on B different bootstrap sample of data using randomly selected subset of features. In this study, for the Random Forest (implemented in WEKA) the number of iteration is set to 100 (*k=100*) and random tree based on the gain ratio is used as its base learner [4].

Support Vector Machine: SVM is considered as the state-of-the-art classification techniques which also attained the best results for PFR [11]. It aims at minimizing the classification error by finding the *Maximal Marginal Hyperplane (MMH)* based on the concept of support vector theory. To find the appropriate support vector, it transforms the input data using the concept of kernel function. In this study, we use SVM with *Sequential Minimal Optimization (SMO)* as a kind of polynomial kernel (implemented in WEKA) which its kernel degree is set to one (*p=1*).

Note that we also used the *Ensemble of Different Classifiers (EDC)* that we proposed in our previous work [2] which attained promising results for similar studies [5, 8]. This classifier consists of five different classifiers (Adaboost.M1, LogitBoost, Naive Bayes, *Multi Layer Perceptron (MLP)*, and SVM) which are trained on the same set of features and combined using majority voting as its algebraic combiner. This classifier is used in this study to evaluated the performance of our proposed approaches. It also used as a tie breaker in the diversely trained ensemble of classifiers proposed in this study.

5 Results and Discussion

To explore the effectiveness of the proposed approaches in this study, we first extract corresponding features to our proposed feature extraction methods for all 55 physicochemical-based attributes explored in this study. Therefore, for a given attribute, a feature group consisting of 21 features is extracted using segmented-based density method and three feature groups consisting of 5, 11, and 21 features are extracted using segmented-based distribution with three different distribution factors explored in this study ($K = 25, 10$ and 5, respectively). We then applied Adaboost.M1, Random Forest, Rotation Forest, and SVM to each feature group. Therefore, for a given attribute, 16 different experiments have been conducted (four classifiers applied to four extracted feature groups).

From the achieve results, we first explore the effectiveness of segmentation factor on the segmented-based distribution method. This experiment is conducted in the following manner. We calculate the average and maximum prediction accuracies achieved for each classifier used in this step with respect to the segmentation factor used in the segmentation-based distribution method for all of the 55 attributes. For example, first we apply SVM to the feature groups extracted from all 55 attributes using segmentation-based distribution method with $K = 5$ separately. Then, we calculate the average and maximum prediction accuracies for all of the 55 achieved results. In the similar manner, SVM is used to feature groups extracted from all 55 attributes using segmentation-based distribution method with $K = 10$ and then for $K = 25$ separately. Then again, we calculate the average and maximum prediction accuracies for all of the 55 achieved results with respect to $K = 10$ and again for all of the 55 achieved results with respect to $K = 25$. In result, 12 maximum and average prediction accuracies are calculated (four average and four maximum prediction accuracies corresponding to three variation of segmentation-based distribution method for SVM, Naive Bayes, AdaBoost.M1, and Random Forest).

In continuation, for a given classifier, we subtract maximum and average values calculated using segmented-based distribution with $K=25$ feature extraction method from the average and maximum values calculated using segmented-based distribution with $K=10$ as well as $K=5$. the results achieved in this step are shown in Table 2. As it is shown in this table, by adding just few features by adjusting segmentation factor from 25% to 5%, for the average, up to 6.9% for EDD dataset and 8.3% for TG dataset prediction enhancements and for the maximum, up to 12.3% for the EDD dataset and 11.5% for the TG dataset prediction enhancements are achieved. Similarly, by adjusting the distribution factor from 25% to 10%, for the average up to 4.8% for EDD dataset and 5.7% for TG dataset prediction enhancements and for the maximum, up to 7.9% for the EDD dataset and 10.2% for the TG dataset prediction enhancements are achieved. These results highlights the effectiveness of our proposed feature extraction methods with respect to the number of extracted features. Note that the performance of Naive Bayes is not improved due to the correlation of the extracted features and therefore is not explored in this part.

Next, we have generate eight different feature sets consisting of combination of features extracted from different attributes using our proposed feature extraction methods in the following two steps. We first study the performance of a given classifier, based on the employed feature extraction method (explored

Table 2. Comparison of the achieved results (%) using Adaboos.M1, Random Forest, and SVM to evaluate the enhancement achieved considering the segmentation-based distribution approach

	EDD		EDD	
AdaBoost.M1	From 25% to 5%	From 25% to 10%	From 25% to 5%	From 25% to 10%
Average	8.3	5.7	6.6	4.3
Maximum	11.3	10.2	10.3	7.9
Random Forest	From 25% to 5%	From 25% to 10%	From 25% to 5%	From 25% to 10%
Average	7.8	5.6	6.9	4.8
Maximum	11.5	9.3	12.2	7.3
SVM	From 25% to 5%	From 25% to 10%	From 25% to 5%	From 25% to 10%
Average	3.8	2.1	3.6	2.1
Maximum	7.1	6.5	6.9	7.1

on the TG dataset). And then, based on each classifier, two feature sets are constructed in the way that each feature set consists of features extracted using similar feature extraction method with the best performances (totally eight combinations). These feature sets have been constructed in the manner to maintain the number of employed features small. In the following paragraph, attributes as well as feature extraction method used to build each of our eight feature sets are explained. For simplicity, we refer to each attribute by its number as in Table 1.

The first and second combinations are extracted respectively based on the performance of the Adaboost.M1 classifier on the segmented-based distribution (with *K=10%*) (attribute numbers: 3, 4, 5, 14, 17, 26, 28, 30, 33, 41, 48 = 121 features) and the segmented-based density (with *K=5%*) feature extraction methods (attributes numbers: 1, 3, 4, 20, 54, 55 = 126 features). The third and forth are extracted based on the performances of the Random Forest classifier on the segmented-based density (with *K=5%*) (1, 3, 16, 17, 41, 55 = 126 features) and the segmented-based distribution (with *K=10%*) (3, 4, 5, 14, 16, 17, 26, 28, 30, 41, 44, 48 = 132 features) feature extraction approaches. The fifth and sixth combinations are extracted based on the performances of the SVM classifier on the segmented-based distribution (with *K=25%*) (1, 3, 4, 5, 17, 27, 29, 30, 31, 33, 35, 37, 38, 39, 40, 41, 44, 47, 48, 55 = 100 features) and the segmented-based distribution (with *K=5%*) (3, 5, 15,17, 30, 41, 44 = 147 features) feature extraction methods. Finally, the seventh and eighth are extracted based on the performances of the Naive Bayes classifier on the segmented-based distribution (with *K=25%*) (1, 3, 4, 5, 14, 16, 17, 27, 29, 30, 31, 32, 33, 37, 38, 39, 40, 41, 44, 47, 48, 55 = 110 features) and the segmented-based density (with *K=5%*) (3, 16, 17, 24, 33, 42 = 126 features) feature extraction methods. It is important to highlight that most of the attributes used to construct these feature sets have not been used or adequately explored for PFR. However, these attributes individually outperform most of the popular attributes used to tackle this problem (e.g. average long range contact energy (16), total non bounded energy (22), and mean fractional area loss (51)).

In continuation, composition of the amino acid feature group (the percentage of occurrence of the amino acids along the protein sequence divided by the length of proteins) as well as the length of the amino acids feature (which attained

Table 3. Results achieved (in percentage %) by using AdaBoost.M1 (Ada), Random Forest (RF), Naive Bayes (NA), SVM, and EDC for 15 feature vectors extracted from the combination of features are extracted in this step (for both EDD and TG datasets).

Datasets	TG					EDD				
Comb_Numb	Na	SVM	Ada	RF	EDC	Na	SVM	Ada	RF	EDC
Comb_1	32.0	40.5	39.1	35.5	**42.7**	38.3	47.4	46.3	41.4	**50.0**
Comb_2	34.1	36.0	38.2	35.1	41.2	39.8	44.6	44.4	38.6	49.2
Comb_3	32.2	36.8	39.0	34.4	41.3	39.4	43.9	45.3	39.7	49.2
Comb_4	34.9	39.3	38.5	37.7	42.4	38.0	44.1	46.2	41.5	47.0
Comb_5	32.7	37.9	38.6	37.9	40.5	37.2	45.0	44.9	42.6	46.4
Comb_6	28.4	40.9	36.9	33.0	39.8	35.8	47.9	44.0	41.6	49.8
Comb_7	30.8	38.3	39.1	36.1	41.2	37.4	44.7	45.1	40.8	47.5
Comb_8	33.0	34.4	37.5	33.6	38.5	42.5	44.2	45.2	38.9	48.6

good results in previous studies [3]) are added (20 + 1 features in total) to each extracted combination of feature groups (which for the rest of this study will be referred as comb_1 to comb_8 respectively). We then apply the employed classifiers in this study to each combination. The results are shown in Table 3. We also apply the *Ensemble of Different Classifiers (EDC)* proposed in [2] which attained the best results for similar studies found in the literature to the extracted combination of features. To compare our results with previous studies, we reproduce the results achieves using EDC to the features extracted in [8] (219 features), extracted features in [5] (125 features), and the 69D feature vector (the 49D feature vector extracted in [7] in addition to the composition of the amino acid feature group (49 + 20 = 69 features)). By reproducing this results we respectively achieve to 48.8%, 47.6%, and 40.7% prediction accuracies for the EDD dataset and 41.1%, 40.7%, and 33.0% for the TG dataset.

As it is shown in Table 3, by using EDC to Comb_1 we achieve to 50.0% and 42.7% prediction accuracy, up to 1.2% and 1.6% better than the best results reported in the literature for similar studies. These results are emphasize on the effectiveness of using features extracted from a wide range of physicochemical-based attributes. To explore even a further range of physicochemical-based attributes with respect to our proposed feature extraction methods, we propose *Ensemble of Diversely Trained AdaBoost.M1 and SVM Classifiers (EDTAS)* in the following manner. We first train two AdaBoost.M1 classifiers diversely trained with Comb_1 and Comb_3 feature vectors and two SVM classifiers diversely trained with Comb_1 and Comb_6 feature vectors. Then for a given test sample, we produce the output of the system using EDC classifier which is trained with Comb_1 feature vector as a tie breaker for two different cases. In the first case, when a fold reached to majority of the votes, it will be directly chosen as the output which out consideration of the EDC classifier. While, in case that a fold would not reach to the majority (two fold with two votes or four different folds with one vote each), the output of EDC will be directly chosen as the output of the system. The architecture of the EDTAS is shown in Figure 2.

Using EDTAS, we achieve up to 50.9% and 43.5% prediction accuracies, up to 2.1% and 2.4% better than previously reported results for the similar studies for the EDD and TG datasets, respectively. The results achieved in this study com-

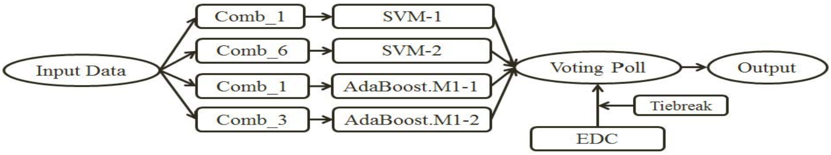


Fig. 1. The overall architecture of the EDTAS

pared to previous results found in the literature for similar studies are reported in Table 4. We also achieved up to 14.3% and 10.5% better prediction accuracy than using 69D feature vector which emphasize on the effectiveness of our proposed feature extraction methods to reveal more discriminatory information from a wide range of physicochemical-based attributes compared to previously used approaches found in the literature [7].

Table 4. The best results (in percentage %) achieved in this study compared to the best results found in the literature for the EDD and TG benchmarks respectively

Study	Attributes (Number of features)	Method	EDD (Results)	TG (Results)
[5]	Features proposed in [5] (126)	SVM	46.3	38.5
[3]	Features proposed in [5] (126)	Ada	44.7	36.4
[4]	Features proposed in [5] (126)	RF	42.9	37.1
[2]	Features proposed in [5] (126)	EDC	47.6	40.7
[8]	Features proposed in [8] (219)	SVM	47.3	40.1
[3]	Features proposed in [8] (219)	Ada	45.3	37.2
[4]	Features proposed in [8] (219)	RF	43.9	38.1
[2]	Features proposed in [8] (219)	EDC	48.8	41.1
[7]	69D (49+20)	SVM	36.6	33.0
This study	Comb_1 (202)	EDC	50.0	42.7
This study	Comb_2 (202)	EDC	49.2	41.3
This study	Comb_3 (202)	EDC	49.2	41.2
This study	Fused (202 for each classifier)	EDTAS	**50.9**	**43.5**

6 Conclusion

In this study we proposed two novel feature extraction methods namely segmented-based density and segmented-based distribution to reveal more local discriminatory information compared to similar approaches found in the literature. We also explored the effectiveness of 55 different physicochemical-based attributes that mostly have not been studied adequately for PFR. We evaluated our proposed approaches using five different classification techniques namely, Naive Bayes, Random Forest, AdaBoost.M1, SVM, and EDC. Then, we generate eight different combination of features extracted from a wide range of attributes based on the results of previous step. Finally, by proposing *Ensemble of Diversely Trained Adaboost.M1 and SVM (EDTAS)* we enhanced the protein fold prediction accuracy for more than 2% better than previously reported results for the similar studies found in the literature.

References

1. Kavousi, K., Moshiri, B., Sadeghi, M., Araabi, B.N., Moosavi-Movahedi, A.A.: A protein fold classifier formed by fusing different modes of pseudo amino acid composition via pssm. Computational Biology and Chemistry 35(1), 1–9 (2011)
2. Dehzangi, A., Phon Amnuaisuk, S., Ng, K.H., Mohandesi, E.: Protein Fold Prediction Problem Using Ensemble of Classifiers. In: Leung, C.S., Lee, M., Chan, J.H. (eds.) ICONIP 2009, Part II. LNCS, vol. 5864, pp. 503–511. Springer, Heidelberg (2009)
3. Krishnaraj, Y., Reddy, C.K.: Boosting methods for protein fold recognition: An empirical comparison. In: Proceedings of the 2008 IEEE International Conference on Bioinformatics and Biomedicine, pp. 393–396 (2008)
4. Dehzangi, A., Phon-Amnuaisuk, S., Dehzangi, O.: Using random forest for protein fold prediction problem: An empirical study. Journal of Information Science and Engineering 26(6), 1941–1956 (2010)
5. Ding, C., Dubchak, I.: Multi-class protein fold recognition using support vector machines and neural networks. Bioinformatics 17, 349–358 (2001)
6. Kurgan, L.A., Zhang, T., Zhang, H., Shen, S., Ruan, J.: Secondary structure-based assignment of the protein structural classes. Amino Acids 35, 551–564 (2008)
7. Gromiha, M.M., Oobatake, M., Sarai, A.: Important amino acid properties for enhanced thermostability from mesophilic to thermophilic proteins. Biophysical Chemistry 82, 51–67 (1999)
8. Dehzangi, A., Phon-Amnuaisuk, S.: Fold prediction problem: The application of new physical and physicochemical- based features. Protein and Peptide Letters 18(2), 174–185 (2011)
9. Chen, K., Kurgan, L.A.: Pfres: protein fold classification by using evolutionary information and predicted secondary structure. Bioinformatics 23(21), 2843–2850 (2007)
10. Shen, H.B., Chou, K.C.: Predicting protein fold pattern with functional domain and sequential evolution information. Journal of Theoretical Biology 256(3), 441–446 (2009)
11. Yang, J.Y., Chen, X.: Improving taxonomy-based protein fold recognition by using global and local features. Protein 79(7), 2053–2064 (2011)
12. Shen, H.B., Chou, K.C.: Ensemble classifier for protein fold pattern recognition. Bioinformatics 22, 1717–1722 (2006)
13. Taguchi, Y.H., Gromiha, M.M.: Application of amino acid occurrence for discriminating different folding types of globular proteins. BMC Bioinformatics 8(1), 404 (2007)
14. Mathura, V.S., Kolippakkam, D.: Apdbase: Amino acid physico-chemical properties database. Bioinformation 12(1), 2–4 (2005)
15. Freund, Y., Schapire, R.E.: Experiments with a new boosting algorithm. In: International Conference on Machine Learning, pp. 148–156 (1996)
16. Witten, I., Frank, E.: Data Mining: Practical Machine Learning Tools and Techniques, 2nd edn. Morgan Kaufmann, San Francisco (2005)
17. Breiman, L.: Random forests. Machine Learning 45, 5–32 (2001)

Realtime Pointing Gesture Recognition and Applications in Multi-user Interaction

Hoang-An Le[1], Khoi-Nguyen C. Mac[1], Truong-An Pham[1],
and Minh-Triet Tran[2]

[1] Advanced Program in Computer Science, University of Science, HCMC, Vietnam
{lhan,mcknguyen,ptan}@apcs.vn
[2] Faculty of Information Technology, University of Science, HCMC, Vietnam
tmtriet@fit.hcmus.edu.vn

Abstract. Pointing is a common gesture of human. Indeed, people tend to involve pointing action in their daily activities using not only bare hand but also with gloves or a kind of pointers like pens, rulers, long sticks, batons, etc. Thus, in this article, the authors propose a new concept of interaction that is centered by the natural gesture of human and a method to detect it under various circumstances. Different from some common approaches which rely on predefined skin color or markers, the proposed method can segment and detect any pointer tip and allow multiple objects to be processed at a time. The method can run with the average accuracy of 91.0%. In case of multiple users using different pointing objects, the accuracy is slightly reduced to 87.9%. The running time is at most 17.14 ms for 9 objects being processed in parallel, and thus can be applied for real time constraints.

Keywords: pointing, gesture, human computer interaction, natural interaction.

1 Introduction

Human-Computer Interaction (HCI) provides natural ways of communication between humans and computers. This means proposed methods should be as similar to the way people communicate as possible. This motivates the development of methods such as speech recognition (to understand human natural language) [9], eye tracking (to understand human's facial expression) [15], or gesture and action recognition (to understand human body languages) [1], etc.

The popularity of pointing action in human's daily communication inspires the authors to propose a new idea of interaction method that based on the concept of human's pointing action. The interaction involves the pointing gesture which has different meanings in different contexts.

This problem is a topic of human gesture recognition whose common approach is to (1) analyze sequences of captured images, (2) detect hand and finger portions, and (3) classify the sequences to some categories that have similar semantic. The classification is based on machine learning method and does not require

A. Selamat et al. (Eds.): ACIIDS 2013, Part I, LNAI 7802, pp. 355–364, 2013.

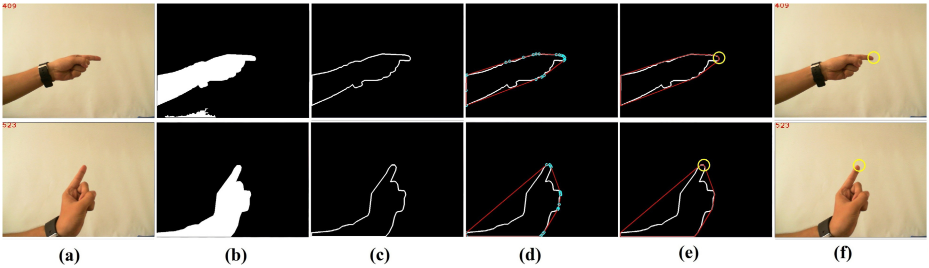
insight understanding of the gestures' special properties which if being exploited, lead to a simpler but efficient solution. Besides, the detection of hand and finger is based on the assumption of predefined values such as skin color [3, 10, 2] or trained markers [8, 14]. Since people tend to use extra objects, or pointers (police's signal light batons, presenters' sticks, etc.) to point at something that they cannot reach, these methods reduce the naturalness of the interaction (as they require users to wear special markers or perform preliminary configuration) and thus become inefficient [6].

As observed from human-human communication, the pointing actions consists of 3 special characteristics: (1) the sharpness (level of protrusion of pointing tips), (2) the first-appeared point (time instant that a tip is detected), and (3) the farthest point (distance from a pointing tip to the frame edge that it appears.) Exploiting the three characteristics, the authors propose a lightweight method that supports all means of pointing interaction regardless of the kinds the objects should be. Instead of using color and trying to detect the pointing tip in every captured frame, the proposed method based on the special geometry shape of pointing objects, i.e. protrusion, to detect the tips and tracking them through frames series, and thus can be performed in real time.

2 Related Work

There are several approaches in HCI such as developing hardware devices like keyboards, mice, cyber-gloves, magnetic tracking devices, etc., or implementing environment-dependable systems using computer vision such as skin color, fingertips or full-body detection, etc. In the scope of this paper, the authors focus only on hand-finger based interactions which provides natural connection between human and computers [4, 7, 13, 5]. In such systems, there are two kinds of approach: sensing based and vision based. The vision-based approach requires a single or multiple cameras, color or infrared, and color gloves or markers. This approach's accuracy, yet, has some drawbacks since it depends on lighting condition and hard to be maintained for sudden changes of surrounding environment [5]. The sensing-based approach, which gives robust performance because of using electronic devices, is limited in only touch interaction [5].

Hand-finger tracking is a noticeable topic of HCI. The goal of tracking algorithms is to predict and estimate the object's position. The approach requires several constraints in order to achieve high performance [17]. For color segmentation-based methods, the performance depends on the color of wearing gloves; for wave-let based methods, the computational cost is expensive and cannot be used for real-time application; for contour based methods, it requires restricted backgrounds to run; for infrared segmentation, the availability of expensive infrared cameras is needed; for correlation-based methods, it is necessary to explicitly setup the stage beforehand; for blob-model based method, the constraint is about the maximum speed of hands' movement [17].



(a) (b) (c) (d) (e) (f)

Fig. 1. Pointing tip detection method. (a) The pointing tip; (b) segmentation phase: a foreground mask of the pointing object; (c) pointing object contour; (d) the convex hull of the object contour with points passed through protrusion level k; (e) the farthest point among points in convex hull; (f): the pointing tip is detected.

3 Proposed Methods

Pointing action is popular in human's activities because it helps to identify objects, get audience's focus, or within different contexts, to select things, direct ways, emphasize, etc. To overcome the limitation of other methods that based on assumptions of skin colors, hand shapes or markers, this article detects pointing actions based on the pointing objects' silhouettes, level of protrusion, and distance to image frame's boundary. As observed, the desired points are the farthest points in an image frame because people tend to stretch their arms to reach or point at something (it is unlikely to point at something by just slightly extending the hands.) In addition, the points should satisfy the sharpness constraint as people usually choose pointing tips that stick out of an object and have substantial level of protrusion like a stick, a pen tip, etc.

The proposed method is based on the special geometrical shapes of pointing tips, which are protrusions of objects such as fingertips, pen tips, etc. The method consists of two steps: segmentation (Figure 1b) and identification (Figure 1c, d, e.) The pointing object is extracted from a captured image using background subtraction method called Codebook [11]. The pointing tip is detected from the segmentation based on its level of protrusion [16]. To reduce the cost of computation, not all points in the object's contour are chosen to computed but only the points that reach into the image (computed from the frame margin it had appeared) farther than a predefined threshold (Figure 1c, d, e).

Instead of trying to detect the pointing tip through every frame which may lead to low performance and error-prone, the authors propose to use the Kalman filter [18] to calculate the optimal coordinates of the detected point based on previous results.

3.1 Pointing Objects Segmentation

The authors use background subtraction to extract pointing objects from the other image parts which plays as a stable background. However, as the background is not exactly a constant image but contains two parts: small static environmental regions and a screen portions which are changed when users trigger an event. On this manner, the background subtraction is carried out not only at the beginning but also right after an event is generated by users to re-train the background model. It should be noticed that the system does not need to continuously update the background but only when the system accepts an event and have the visual content changed.

Capable for both illumination change and moving-background training, the authors propose to use the Codebook algorithm [11]. When re-training the background, some portion of the new background may be occluded by a user's pointing object. To solve this issue, the authors correct the occluded area in the background by applying the homographic transform on the current screen content and mapping the corresponding region into the background.

The background subtraction phase runs through the following steps

- Periodically study a model of the background. During background learning, if a codeword is not accessed for a period of time, it is deleted and replaced by nonstale, i.e. active entries.
- Obtain foreground objects by using the learned model to segment it out of the background.
- Periodically clean out stale codebook entries and update the learned background pixels after a period of time.

3.2 Pointing Tip Detection

The purpose of this section is to identify the hot spot of each pointing object detected from previous phase and the location where users are pointing at. For each pointing object, the tip detection runs through 4 consecutive steps:

- Firsly, because a hot spot can only occur at the boundary an object the authors have the contour of each pointing object extracted.
- The set of points that forms the convex hull of the contour is then computed. Since the pointing tip is usually the part that sticks out of an object, this set is the candidate points for the desired hot spot.
- For each point in the candidate set, the authors compute the level of protrusion and only those that pass a predefined threshold can be present.
- Finally, among the candidate, the hot spot is selected to be the farthest point computed from the frame edge that it first appears.
- In order to improve the accuracy, the authors apply the Kalman filter that uses the previous knowledge of the hot spot to correct the new detected point from noises that may appear.

Contour Extraction. Let $\mathcal{M}$ be the foreground mask obtained from any arbitrary frame after the segmentation phase, $\mathcal{B} \in \mathbb{N}^2$ be a blob or a connected component in $\mathcal{M}$ and $\mathcal{C} = \beta(\mathcal{B})$ be the contour of the blob $\mathcal{B}$,

$$\mathcal{C} = \beta(\mathcal{B}) = \mathcal{B} - (\mathcal{B} \ominus K),$$

where $\mathcal{B} \ominus K$ is the erosion of $\mathcal{B}$ and a 3×3 matrix K, which consists of all points p such that K translated by p, is contained in $\mathcal{B}$, i.e. $\mathcal{B} \ominus K = \{p | (K)_p \subseteq \mathcal{B}\}$. The result of each pointing object in Figure 1b is shown corresponding in Figure 1c.

Convex Hull Computation. As the matter of fact, it is possible to find the pointing tip right after the object contour is extracted. However, since the contour may become zigzag due to noises from background subtraction Figure 2a, b, or complicated Figure 2c because of the shape of the pointing object used, the convex hulls $\mathcal{H}$ of the countour $\mathcal{C}$, $\mathcal{H} \subset \mathcal{C}$ is computed.

Since the pointing tip is a protrusion part on an object's contour, the convex hull computation does not leave out the correct tip (see Figure 1d, e and the third column of Figure 2) but only narrowed down the number of points to be checked. Thus, the step helps to increase the system performance.

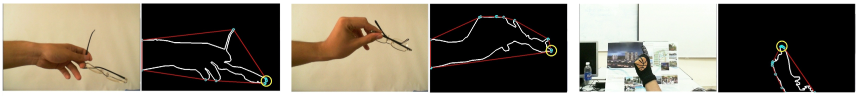


Fig. 2. Convex Hull

Level of Protrusion. Based on the sharpness level k of each point, the candidate set is filtered to those that are sharper than a threshold level k_{thr}. The sharpness level k of a point is computed based on the angle it makes with 2 arbitrary points in the neighborhood:

$$k_i = p_1^- p_2^+ - p_2^- p_1^+$$

For each point $P_i \in \mathcal{H}$ in the convex hull of the contour, let $P_{i-k}, P_{i+k} \in \mathcal{C}$ be any two points on the contour, may or may not be in the convex hull. The authors have $\mathbf{p}^-$ and $\mathbf{p}^+$ are 2 vectors defined by:

$$\mathbf{p}^- = (p_1^-, p_2^-) = \overrightarrow{P_{i-k}P_i}, \text{ and } \mathbf{p}^+ = (p_1^+, p_2^+) = \overrightarrow{P_i P_{i+k}}$$

The threshold level k_{thr} is different for each kind of objects such as the threshold level of a pen-tip is smaller than that of a finger, etc.

Pointing Tip Detection. Let $\mathcal{S}_C$ be the candidate set narrowed down from the convex hull of the contour, i.e. $\mathcal{S}_C = \{P_i | P_i \in \mathcal{H}, k_i > k_{thr}\}$. Since people tend to stretch their arm to reach for the pointed objects, the pointing tips should be the farthest points among all the candidates. By farthest, the authors mean that the distant d of each candidate point is computed from the frame edge that the object first appears. For instance, candidate points on the object in the first row of Figure1 are measured to the left boundary whereas candidate points on the second and third row are measured to the bottom boundary.

Correction. Because of sudden changes in environmental illumination or occlusions, noises and errors may happen during the process. To bypass the wrong detection, the authors propose using a method to correct the detection result. By observation, within a short period of $10 - 500ms$ users usually have the pointing objects move in a stable direction. Thus, the Kalman Filter is selected to do the correction step. Using the Filter, the authors track the pointing tip from the first time it appears in the viewport, and in every step, the optimal point is computed from both the measurement, i.e. the newly detected result, and the result from previous steps.

4 Experiments and Results

4.1 Dataset Description

The test scenarios contain sets of video clips which are divided into 3 groups A, B, and C:

- Group A includes rigid objects whose shapes do not change much: bare hand ($A1$), finger ($A2$), gloved hand ($A3$), gloved finger ($A4$), and pen ($A5$.)
- Group B includes deformable objects whose shapes change significantly: deformable hand ($B1$), deformable finger ($B2$), glasses ($B3$), blinking semi-transparent stick ($B4$), and slider ($B5$.)
- Group C includes multiple hands ($C1$), multiple pens ($C2$), blinking semi-transparent stick and slider ($C3$), 3 users with different pointing objects ($C4$), and 5 users with different pointing objects ($C5$.)

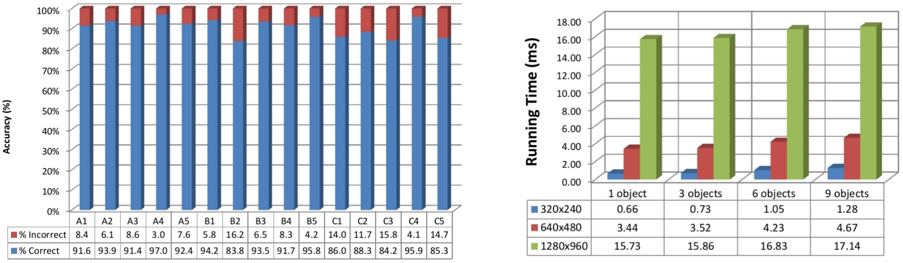
The test video clips are recorded with $25 frames/ms$; each is 1000-frames long of 3 different resolutions 320×240, 640×480, and 1280×960 pixels. The pointing objects are chosen so that their colors are visually recognizable from the background. In addition, the test cases vary from single object to multiple ones to check the performance and accuracy under different conditions. Besides, the objects' colors and shapes are also picked randomly to guarantee that there is no predefined color or shape used in the test scenarios.

4.2 System Accuracy

The system accuracy, is calculated from the number of frames (out of 1000) giving correct results (Figure 3a.) By experiment, the accuracy over group 1 is 1.4% greater than group 2 and the accuracy over group 2 is 3.9% greater than group 3. It means that, among the 3 groups of test cases, rigid objects are the easiest one to processed while multiple objects used by several users are, on the other hand, the most difficult ones. The difference, however, is not significant as the complexity over the test scenarios increases. Therefore, the system is able to produce high accuracy (with mean of 91.0%) which is stable over several test cases (with standard deviation of 4.4%) and can be applied in real life situations.

4.3 System Performance

System performance is measured using the total running time of the test cases classified by the number of objects: 1, 3, 6, and 9 (Figure 3b). As the frames' resolutions increase, the running times also increase with the mean duration of 3.03ms over the first 2 resolutions, and 12.4ms over the last 2 resolutions. It means that the growth of frames' resolution increases the running time. Besides, the mean running times for the 3 resolutions (over the 4 test cases) are respectively 0.93ms, 3.96ms, and 16.39ms. Therefore the processing time is not affected by the screen's resolution but by the number of objects.



	A1	A2	A3	A4	A5	B1	B2	B3	B4	B5	C1	C2	C3	C4	C5
🟥 % Incorrect	8.4	6.1	8.6	3.0	7.6	5.8	16.2	6.5	8.3	4.2	14.0	11.7	15.8	4.1	14.7
🟦 % Correct	91.6	93.9	91.4	97.0	92.4	94.2	83.8	93.5	91.7	95.8	86.0	88.3	84.2	95.9	85.3

(a)

	1 object	3 objects	6 objects	9 objects
🟦 320x240	0.66	0.73	1.05	1.28
🟥 640x480	3.44	3.52	4.23	4.67
🟩 1280x960	15.73	15.86	16.83	17.14

(b)

Fig. 3. Accuracy (a) and system performance (b) of different pointing objects

As the number of pointing objects increases, the processing time also increases. However, the two growth levels do not depend on each other because the processing phase for each object contains similar steps (background subtraction and contour extraction,) which are processed at the same time. Therefore, the processing time only increase by an insignificant amount as the number of objects increases.

On the other hand, the running time depends greatly on the frame resolution as the time is proportional to the number of pixels in a frame, i.e. when the resolution is doubled, the running time increases around 4 times. For the test scenario with the highest resolution, it requires 17.14ms at most. Hence, the system still satisfies real-time performance criteria in experiment's worst cases.

4.4 Discussion

Figure 4 shows some experiments of bare hands and fingers with their silhouettes. In these experiments, the hands and fingers, are kept rigid as they are moving. It appears that the detected regions match the one the users pointing to.

As an example of deformable objects, the authors use glasses and change their shapes so that the protrusive and farthest points are changed over time (Figure 5). Without using any tracking methods, the system detects new points with higher level of protrusion and lose focus on the original ones, which should be continued to be detected.

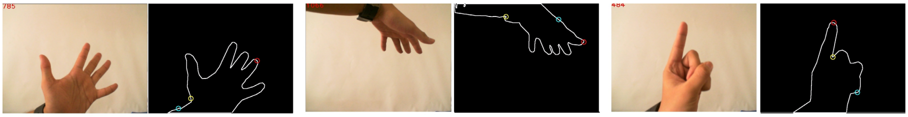


Fig. 4. Hands' silhouettes with detected points (red circle)

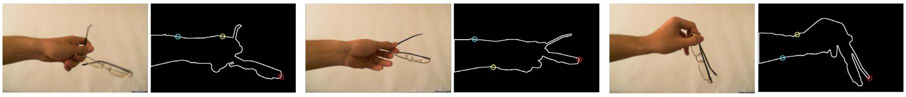


Fig. 5. Deformable glasses' silhouettes with detected points (red circle)

Another case that leads to system inaccuracy is shown in Figure 6 that describes the situation of two objects occluding each other. If the filter is not present, the overlapped point is lost after the occlusion; otherwise, the covered regions are remembered and continued to be tracked.

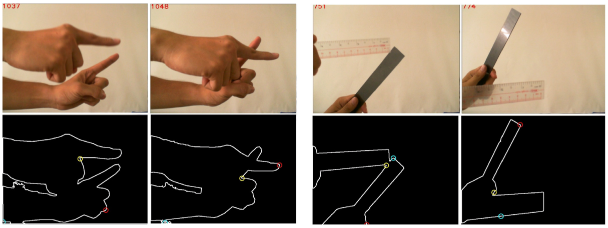


Fig. 6. Occlusion in multiple objects: fingers and rulers

5 Application

The Smart Interactive Map (SIM) [12] transforms normal physical map into a system that can understand users' pointing gestures and offer them information such as the best route, tourist attraction, restaurant, etc (Figure 7.) The system consists of several map stations at different locations in a certain area, each which connects to one processor, for parallel processing. It is designed to serve one user at a time to avoid occlusion caused by multiple users. Instead, two hands of a user are the cause of occlusion, but the frequency of this is low.

Although the system does not serve groups of users, it still shows the parallel processing property of the authors' proposed method. Beside the parallel processing property of the system, understanding pointing gesture of users' fingers or protrusive objects is the most important part of this system. By setting up SIM, the authors show two strong properties of the proposed method that are parallel processing performance and natural way in interaction.

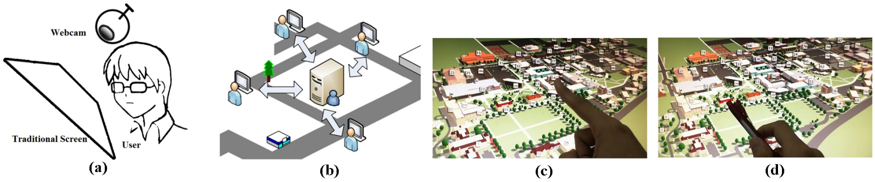


Fig. 7. Smart Interactive Map

6 Conclusion and Future Work

This paper introduces a new kind of interaction method: pointing gesture, to provide a high level of flexibility and naturalness, as it is commonly used. The proposed method can overcome existing obstacles of using predefined colors and shapes. It allows people to use any arbitrary objects to interact with the systems without learning their special features since the method depends on the protrusive regions of the objects.

By experiment, the system takes 17.14ms with an accuracy of 91.0% to process 9 objects in parallel. The error is due to lost tracking when the objects become non-rigid or when occlusion happens. As people tend to move pointing objects slowly and linearly in short periods of time, the author propose to use Kalman filter to further improve the method's accuracy.

This method can be used as a low cost replacement for current interactive systems. For example, the method can be applied to build (1) interactive map guiding systems in large areas, such as campuses, amusement parks, shopping malls, etc; (2) teaching or presenting systems, where users have to use pointing gestures most of the time; (3) entertainment systems which requires high flexibility while operating.

Acknowledgement. This research is supported by research funding from Advanced Program in Computer Science, University of Science, Vietnam National University - Ho Chi Minh City.

References

1. Aggarwal, J., Ryoo, M.: Human activity analysis: A review. ACM Comput. Surv. 43(3), 16:1–16:43 (2011), `http://doi.acm.org/10.1145/1922649.1922653`
2. Choi, S.-H., Han, J.-H., Kim, J.-H.: 3D-Position Estimation for Hand Gesture Interface Using a Single Camera. In: Jacko, J.A. (ed.) HCI International 2011, Part II. LNCS, vol. 6762, pp. 231–237. Springer, Heidelberg (2011)
3. Dawod, A.Y., Abdullah, J., Alam, M.J.: Fingertips detection from color image with complex background. In: The 3rd International Conference on Machine Vision, ICMV 2010, pp. 88–96 (2010)
4. Dietz, P., Leigh, D.: DiamondTouch: a multi-user touch technology. In: Proceedings of the 14th Annual ACM Symposium on User Interface Software and Technology, UIST 2001, pp. 219–226. ACM, New York (2001)

5. Do-Lenh, S., Kaplan, F., Sharma, A., Dillenbourg, P.: Multi-finger interactions with papers on augmented tabletops. In: TEI 2009: Proceedings of the 3rd International Conference on Tangible and Embedded Interaction, pp. 267–274. ACM, New York (2009)
6. Fernandes, B., Fernández, J.: Bare hand interaction in tabletop augmented reality. In: SIGGRAPH 2009: Posters, pp. 98:1–98:1. ACM, New York (2009)
7. Han, J.Y.: Low-cost multi-touch sensing through frustrated total internal reflection. In: Proceedings of the 18th Annual ACM Symposium on User Interface Software and Technology, UIST 2005, pp. 115–118. ACM, New York (2005)
8. Jones, M.J., Rehg, J.M.: Statistical color models with application to skin detection. International Journal of Computer Vision 46(1), 81–96 (2002)
9. Juang, B.H., Rabiner, L.R.: Automatic speech recognition - a brief history of the technology development. Elsevier Encyclopedia of Language and Linguistics (2005)
10. Kang, S.K., Nam, M.Y., Rhee, P.K.: Color based hand and finger detection technology for user interaction. In: Proceedings of the 2008 International Conference on Convergence and Hybrid Information Technology, pp. 229–236. IEEE Computer Society, Washington, DC (2008)
11. Kim, K., Chalidabhongse, T.H., Harwood, D., Davis, L.: Background modeling and subtraction by codebook construction, vol. 5, pp. 3061–3064 (2004)
12. Le, H.A., Mac, K.N.C., Pham, T.A., Nguyen, V.T., Tran, M.T., Duong, A.D.: SIM - smart interactive map with pointing gestures. In: 4th International Conference on Intelligent Human-Machine Systems and Cybernetics, IHMSC 2012, vol. 2, pp. 344–349 (2012)
13. Letessier, J., Bérard, F.: Visual tracking of bare fingers for interactive surfaces. In: Proceedings of the 17th Annual ACM Symposium on User Interface Software and Technology, UIST 2004, pp. 119–122. ACM, New York (2004)
14. Mistry, P., Maes, P., Chang, L.: Wuw - wear ur world: a wearable gestural interface. In: Proceedings of the 27th International Conference Extended Abstracts on Human Factors in Computing Systems, CHI EA 2009, pp. 4111–4116. ACM, New York (2009)
15. Model, D., Eizenman, M.: User-calibration-free remote gaze estimation system. In: ETRA, pp. 29–36 (2010)
16. Segen, J., Kumar, S.: Shadow gestures: 3D hand pose estimation using a single camera. In: IEEE Computer Society Conference on Computer Vision and Pattern Recognition, vol. 1, p. 485 (1999)
17. Song, P., Winkler, S., Gilani, S.O., Zhou, Z.: Vision-Based Projected Tabletop Interface for Finger Interactions. In: Lew, M., Sebe, N., Huang, T.S., Bakker, E.M. (eds.) HCI 2007. LNCS, vol. 4796, pp. 49–58. Springer, Heidelberg (2007)
18. Welch, G., Bishop, G.: An introduction to the kalman filter (1995)

Multi-domain Public Key Infrastructure for Information Security with Use of a Multi-Agent System

Nilar Aye[1], Hlaing Su Khin[1], Toe Toe Win[1], Tayzar KoKo[1], MoMo Zin Than[1], Fumio Hattori[2], and Kazuhiro Kuwabara[2]

[1] Yatanarpon Teleport Co., Ltd, Universities' Hlaing Campus, Yangon, Myanmar
{nilaraye,hlaingsukhin,toetoewin,tayzarkoko,
momozinthan}@teleport.net.mm
[2] College of Information Science and Engineering, Ritsumeikan University
1-1-1 Noji Higashi, Kusatsu, Shiga-ken 525-8577 Japan
{fhattori,kuwabara}@is.ritsumei.ac.jp

Abstract. We propose a multi-agent based common framework for interoperability among CAs (Certificate Authorities) from multiple PKI (Public Key Infrastructure) domains to build trust and secure for electronic transactions. Most of the countries recognize PKI as a powerful technique for security services and implemented their own PKI for online user. Several trust models have been used in PKI, and achieving interoperability between them is a major issue which requires recognition of certificates from different domains in order to perform transactions confidently in a cross border application. In our system, User Agent, Trust Agent and Management Agent are created. These software agents co-operate each other for user authentication and authorization processes autonomously in multiple PKI domains within the ASEAN region to encourage the recognition of digital signature to enhance regional market. Our system intended to facilitate not only for matured e-commerce but also for individual start up entrepreneur for secure trading by taking into consideration of regional needs.

Keywords: Public Key Infrastructure (PKI), Multi-Agent System (MAS), PKI interoperability, Certification Authority (CA), Digital Certificate.

1 Introduction

Nowadays, the development of online services such as e-Government, e-Commerce, e-Procurement are growing tremendously whilst online fraud and misuse of personal information are also increasing by means of various attacks. Although the web site is being configured with user registration and authentication, user can give wrong information in the registration process. Therefore, verification of buyer and seller is required in an e-commerce arena in order to build trust. The technology of Public Key Infrastructure (PKI) [1][2] has been established to solve these issues such as data integrity, confidentiality, user authenticity, and non-repudiation of online users. With the use of digital certificates a user can identify who is a real buyer or seller as well as

A. Selamat et al. (Eds.): ACIIDS 2013, Part I, LNAI 7802, pp. 365–374, 2013.

individual or organization. However, users need checking manually the certificate's validity, whether it has been issued by trusted authority or not, and its status. According to the user's knowledge, this task would be troublesome as certificate checking processes are tedious and complicated. The situation may get worse when the buyer and seller are using digital certificates from different PKI domains. Therefore, we propose a system using PKI and multi-agent technologies to identify a user and authenticate user access to the system to support secured e-commerce. In this approach, software agents autonomously perform checking of the digital certificates' validity, certificate chain and certificate revocation status instead of a user. The major contributions of the system are:

- Strong authentication mechanism based on PKI ;
- Multi-agent cooperation for authentication of user access and validation of digital certificates from various PKI domains; and
- Implement a common framework for CA-CA Interoperability.

This paper is structured as follows. The next section describes the current issues regarding PKI interoperability, and Section 3 discusses some related works that apply the multi-agent concept to PKI. Section 4 presents the motivating scenario for the propose system. Section 5 describes the implementation of the proposed system. Section 6 discusses the merits of the proposed system, and the final section concludes this paper.

2 Current Issues in PKI Interoperability

Nowadays PKI is an important part of information security infrastructure. PKI based e-business security system [3] has been proposed as an effective solution for online commerce. There are different types of digital certificates issued from various CAs which represents an identity of online user. It can be used for a trust relationship and encrypt/decrypt user information for privacy and security. Most countries recognize PKI as a powerful technique for security services and implement their own PKI model for online users. Several trust models have been used in PKI. Analysis and comparison of PKI-based trust model are proposed in [4] to provide information security services. Although the use of certificate has been mutually recognized by each other within a region, achieving PKI interoperability [5] is required for cross border applications to facilitate business trust. Thus, interoperability among different PKI domains becomes an issue in order to promote cross border activities in electronic commerce.

A number of alternatives have been suggested for the PKI interoperability such as Cross-certification, Bridge CA, Certificate Trust Lists, Accreditation Certificate, Strict hierarchy and etc., in [6]. Besides, one of the proposals was presented for achieving inter-domain interoperability [7] and provided recommendations for the way forward. The CA-CA interoperability has been implemented in the European Union following the Trust List model. Although it has been initiated within ASEAN lately, granting mutual recognition will take time due to various constraints such as legal issues, technical standards and so on. Therefore, a method is needed to verify digital certificates in a common application to build trust between users and to secure their information.

Another concern that needs to be taken into consideration is a process of certificate verification and validation by user him/herself who requires knowledge of digital certificate and online security. For instance, user certificate must not be expired, it has not been revoked including its signing certificate, and purpose of key usage must be standard for online applications. There are several steps require for a certificate validation. In order to be a legitimate certificate, for example, a buyer needs checking of a seller's certificate and its issuers' validity period. Besides, issuing authorities must be an authorized trusted third party. Afterwards, the buyer verifies all certificates' revocation status by using online services such as OCSP, LDAP or downloads ARL/CRL manually. These checking processes are complicated requiring autonomous actions which carry out one after another. Therefore, we propose a method to accomplish these issues with the use of a multi-agent system.

3 Related Works

There are a number of research works have been done on combination of PKI and multi-agent technologies. For example, the 'secret agent' concept was proposed, and KQML [8] was extended to handle public key management in [9]. The focus of this paper is on the development of an agent system itself which used KQML as an agent communication language and to make the agent system more secure. So, they introduced PKI into the agent system. In doing so, KQML was extended to include several new performatives to handle digital certificates. Their proposed 'secret agent' handles flexible configuration of digital certificate management. This approach does not prespecify any particular certification format and hierarchical relationship in the software and it is dynamically formed as the agents apply/issue their certificates according to the desires of the applications. But, our proposed system uses pre-defined trusted database which stores certificates and its related hierarchies for verification by the software agent.

Another approach was introduced an agent-oriented public key infrastructure (APKI) for multi-agent e-service [10]. Its application to the digital certificate management was described in [11] by same authors. They argued that their proposed APKI provides a binding mechanism between human and agent so that the legal responsibility of agent can be traced to the corresponding human user. The proposed APKI was built on the FIPA-OS. In this approach, human and agent required certificates for identification which were explicitly separated in order to differentiate between human and agent certificates. Similar to our system, the activities such as authentication, authorization, access control, and trusted relationships are carried out by the use of PKI and multi-agent technologies. However, our proposal requires only user certificate which is used by an agent to identify user and control access of the system after certificate verification.

One more interesting approach also described the concept of multi-agent which is applied to the authentication in the multi-application environment [12]. By introducing an application agent (AA) corresponding to a particular application, the proposed system can flexibly handle the requirements of multiple applications. However, it does not consider the multiple PKI domains. The same authors propose in [13] similar to our system creating several agents for client certificate validation, authorization check, access granting and administration application delegation scheduling. Moreover, it employs PKI to build trust among agents. It is, however, different from our approach in that MAS subscribe to the CA and own the key pair and the certificate that messages among agents can be signed/verified and encrypted/decrypted with the basic PKI scheme. These agents look up the LDAP and verify the authenticity of the client certificate.

4 Motivating Scenario

E-commerce security is a serious issue [14] requiring strong authentication as well as authorization. A solution is needed to control an access of application of both buyer and seller which allows buying and selling can be done at the same place. Products posted on the site can be viewed by everyone. But, user registration is required for a new user and digital certificate is always necessary for whoever wants to engage in an activity such as buying, selling or giving comments to a product. The user can get digital certificate from any trusted third party or Certification Authority (CA) by paying certain fees which must be stored in an e-token, smart card or secured device. Together with a user certificate, its signing certificates and related chain are also included in it. Most of the digital certificates have been issued with one year validity. A system administrator keeps a user database and certificate database up to date and makes it secure by executing necessary steps.

In the motivating scenario, users can perform as a role of buyer or seller who does actions such as uploading product to sell, updating product information, buying something or giving comments. The buyer can identify the product owner easily since a digital certificate has been posted together with the product. When a user clicks to buy, the system automatically carries out verification and required negotiation processes to complete the transaction successfully. Then, a seller receives a notification email from the system and proceeds to the bank for payment confirmation. The user does not need to consider steps of certificates checking. The buyer, seller and bank cooperate with each other to achieve the business processes.

5 Use of Multi Agents

To implement a system that can execute the motivating scenario described above, we apply multi-agent technology. Multi Agent System (MAS) [15] is a technique where several agents communicate each other to solve problems that are difficult or impossible for an individual agent. In our system, User Agent(UA), Trust Agent (TA) and Management Agent(MA) are created which co-operate each other for user authentication and authorization processes (Figure 1).

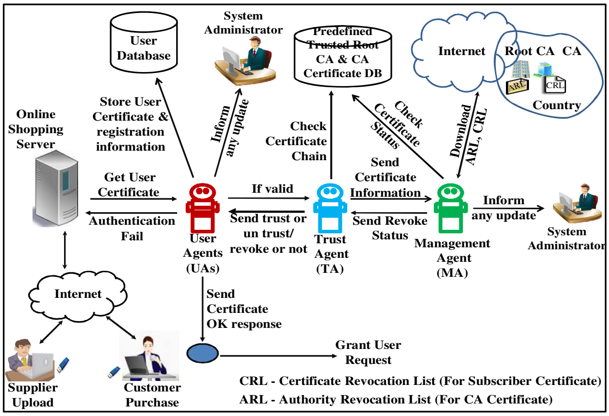


Fig. 1. Motivating Scenario as Implemented as a Multi-Agent System

User Agents are created for every user which is responsible for user authentication to validate the user certificate such as the certificate's public key, its validity and key usage. If the certificate is not valid, it rejects the user to enter the system. Prior to the expiration date, UA sends a message to System Administrator to inform the seller whose certificate will expire soon. If the user fails to update the certificate on time, the UA sends as notification to System Administrator to remove the product from the database. It also sends a message to the TA for revocation checking. The process flow of User Agent is shown in Figure 2.

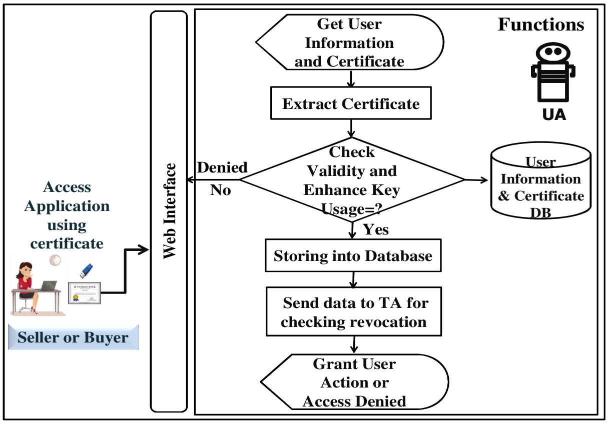

Access Application using certificate
Seller or Buyer
Web Interface
Functions
UA
Get User Information and Certificate
Extract Certificate
Check Validity and Enhance Key Usage=?
Denied
No
Yes
User Information & Certificate DB
Storing into Database
Send data to TA for checking revocation
Grant User Action or Access Denied

Fig. 2. Processes of User Agents

After receiving a message from UA, TA verifies the issuer certificate and its chain based on its trust model in the certificate Database. Suppose that the user certificate is issued from the Hierarchical model. If so, the authority key identifier (AKI) of the user certificate and the subject key identifier (SKI) of its issuer certificate (CA) must match, and AKI of CA certificate and SKI of its issuer certificate (Root CA) are required to check by TA again. After these steps, TA sends a reply message to UA to decline user access or MA to examine the revocation status. The process flow of Trust Agent is shown in Figure 3.

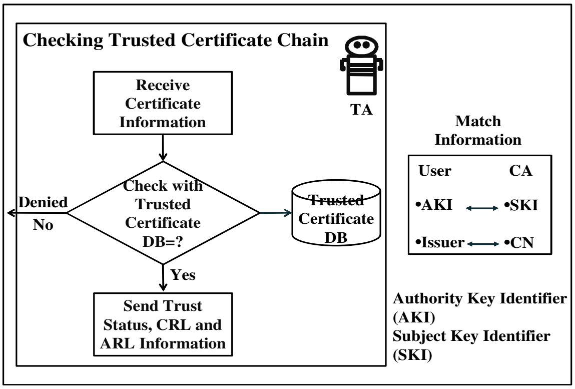

Checking Trusted Certificate Chain
TA
Receive Certificate Information
Check with Trusted Certificate DB=?
Denied
No
Yes
~~Trusted~~ Certificate DB
Send Trust Status, CRL and ARL Information
Match Information
User
CA
•AKI
•SKI
•Issuer
•CN
Authority Key Identifier (AKI)
Subject Key Identifier (SKI)

Fig. 3. The Processes of Trust Agent

MA receives the user certificate information from TA and retrieves Authority/ Certificate Revocation List (ARL/CRL) location from the database. MA downloads a file for checking of certificates. If any of them are included in the revocation list, MA replies to TA not to allow this user to the system. If not, it grants user to perform transaction securely. MA also keeps track of the Trusted Certificate Database for security whether it has been updated or amended. Besides, MA checks validity of certificates frequently and informs status to the administrator by messaging. The process flow of Management Agent is shown in Figure 4.

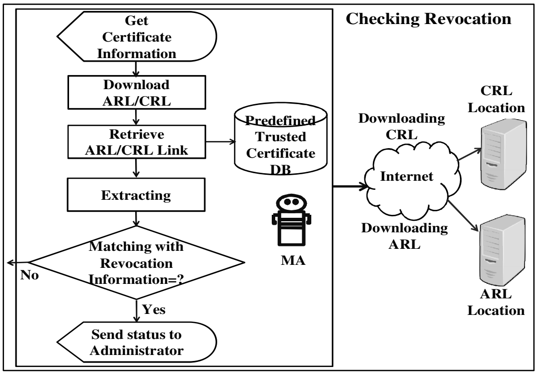


Fig. 4. The Processes of Management Agent

5.1 Messages in Agent Communication

In our proposed approach, we use Agent Communication Language (ACL) to communicate among agents by means of message exchange [16]. The popular ACLs are

Foundation for Intelligent Physical Agents (FIPA) [17] and Knowledge Query and Manipulating Language (KQML). In FIPA, several interaction protocols are defined and we follow its standard and specification for agent implementation. An example agent message exchange is described in Figure 5. In this example, UA extracts certificate into detail and checks validity and related information.

An access of user is denied, if certificate validation process fails. Otherwise, UA sends a “REQUEST” to TA to check the chain of user certificate. The latter checks the related certificate chain with the help of predefined trusted certificate database. For instance, a chain is a series of issuing authority certificates i.e., Root CA and CAs, etc. TA “INFORM”s an error message to UAs, if certificates are not matched with stored certificates which means it is not trusted. If not, TA “ASK”s MA for checking of the revocation status. Then, MA downloads the CRL file from the Internet and checks whether the user certificate or any of the issuing certificates are included in it. MA sends a “REPLY” message with a status of “Revoked or not” to TA. Afterward, TA sends a “REPLY” message to inform UA whether the user access to the system is granted or denied.

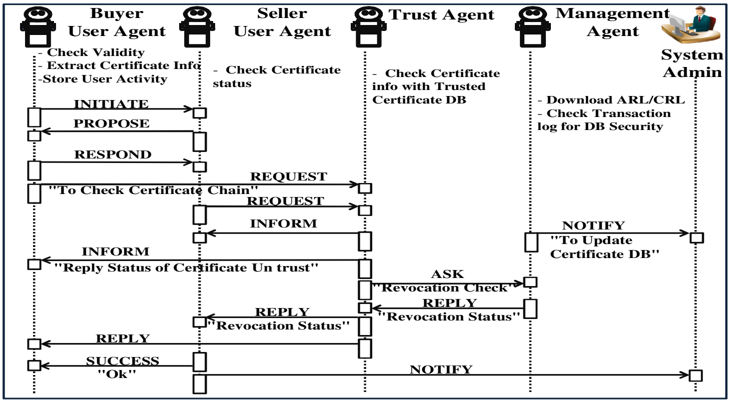
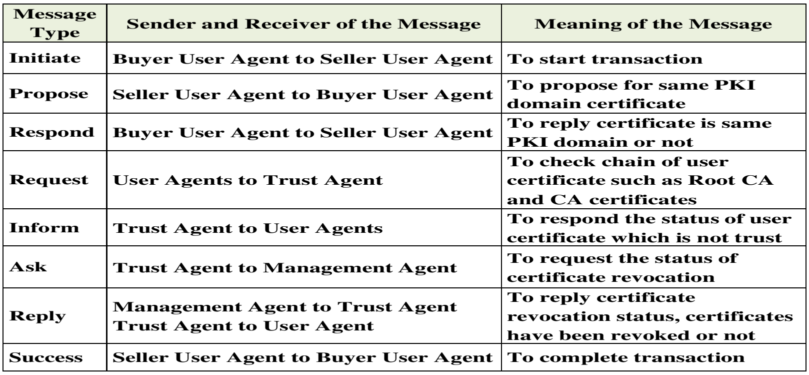


Fig. 5. Agent Communication Steps

The negotiation between user agents to decide on the PKI domain to use can be described as follows (Figure 6):

- UA (Buyer Agent) “INITIATES” with its own certificate to another agent (Seller Agent). If the user certificates are under the same PKI, the negotiation becomes successful and carries out the transaction.
- If not, the Seller Agent “PROPOSE” to Buyer Agent to provide a similar domain certificate
- Buyer Agent “RESPOND” the same PKI certificate to the Seller Agent
- Seller Agent sends “SUCCESS” message to complete the transaction



Message Type	Sender and Receiver of the Message	Meaning of the Message
Initiate	Buyer User Agent to Seller User Agent	To start transaction
Propose	Seller User Agent to Buyer User Agent	To propose for same PKI domain certificate
Respond	Buyer User Agent to Seller User Agent	To reply certificate is same PKI domain or not
Request	User Agents to Trust Agent	To check chain of user certificate such as Root CA and CA certificates
Inform	Trust Agent to User Agents	To respond the status of user certificate which is not trust
Ask	Trust Agent to Management Agent	To request the status of certificate revocation
Reply	Management Agent to Trust Agent Trust Agent to User Agent	To reply certificate revocation status, certificates have been revoked or not
Success	Seller User Agent to Buyer User Agent	To complete transaction

Fig. 6. Summary of the Agents Message

5.2 System Implementation

Java technology is applied to implement the system. We use Apache Tomcat for web service and MySQL for user database and trusted certificate database (Figure 7). They keep user, product, and certificate information. Software agents are created to control user access and activity of the system. Agent automatically downloads ARL/CRL files everywhere from the Internet.

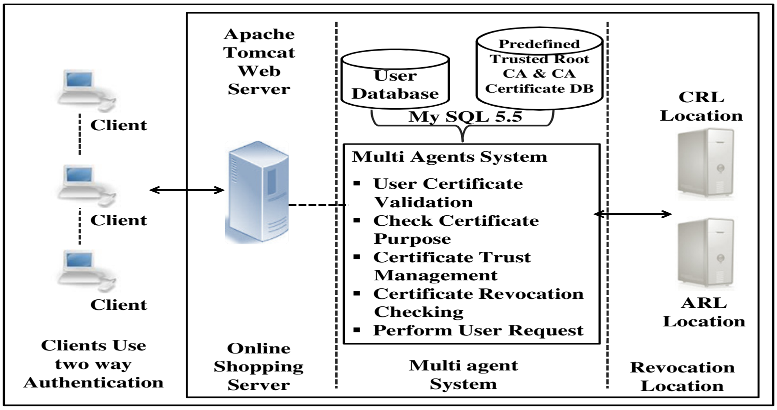

Client
Client
Client
Clients Use two way Authentication
Apache Tomcat Web Server
Online Shopping Server
User Database
Predefined Trusted Root CA & CA Certificate DB
My SQL 5.5
Multi Agents System
User Certificate Validation
Check Certificate Purpose
Certificate Trust Management
Certificate Revocation Checking
Perform User Request
Multi agent System
CRL Location
ARL Location
Revocation Location

Fig. 7. Overview of the system

6 Discussion

Nowadays, the use of PKI becomes complex due to multi-domain and multi-vendor PKI with various implementations. The merit of our system is to ensure the multi-domains PKI interoperability [18] and to contribute the development and spread of reliable PKI in applications of future electronic society. The system will recognize different digital certificates by validating the certificate path and maintain trust entity. The proposed system has the following benefits.

- User-Oriented- The system is easy to be used by everyone from anywhere at any time. Also, it is not only user friendly but also simple from the user point of view which does not require additional hardware and software. Users can get digital certificate within their region with reasonable price.
- Interactivity- PKI based multi-agent system establishes user trust and secure channel within local and international organizations. It can operate in distributed, dynamic and open user environments which make it possible to enrich interactions of the system. The issue of interoperability between CAs can be solved.
- Cost- Saving - Online applications offer saving of time, cost and transportation which are important factors taken into account from the user point of view. Our system introduces secured online transactions to benefit users for doing business regardless of the region and time that can save other expenses such as travelling, accommodation, etc.
- Convenience – Buyer and Seller can communicate as well as do business online easily. Only a digital certificate which is stored in a secured, portable and removable device is required for user authentication and granting access to the system.

7 Conclusion and Future Work

We have presented the idea and implementation of PKI based multi-agent system including verification process of different PKI certificates on behalf of a human user. That solves PKI interoperability issues among different CAs. It also provides trust for online users which facilitates not only online commerce security but also to support startup entrepreneurs for secured trading by taking into consideration regional needs. With the use of our proposed system, user can identify each other easily which can strengthen business trust.

For improvement of our system, we need to consider including all applicable trust models. Currently, we have implemented a Hierarchical trust model as an example model for CA-CA interoperability. In order to apply it in the real world, we require creating a certificate database which must consists of all trust models [19] for the verification processes. Our future work includes the development of an adaptive model for checking certificate path validation of heterogeneous PKI integration to manipulate the shortest and best certification path by using software agents in the changing environments.

Acknowledgements. This research project has been supported by Asia Pacific Telecommunity (APT).

References

1. Adams, C., Lloyd, S.: Understanding PKI: Concept, Standards, and Deployment Considerations, 2nd edn. Addison-Wesley (2002)
2. More, V.N.: Authentication and Authorization Models. International Journal of Computer Science and Security (IJCSS) 5(1), 72–84 (2011)

3. Zhou, H.Q., Dai, S.H.: PKI-based E-Business Security System. In: The 3rd International Conference on Innovative Computing Information and Control, ICICIC 2008 (2008)
4. Liping, H., Lei, S.: Research on Trust Model of PKI. In: Fourth International Conference on Intelligent Computation Technology and Automation, pp. 232–235 (2011)
5. Achieving PKI Interoperability, Results of the JKS-IWG Interoperability project, Japan PKI Forum, Korea PKI Forum, PKI Forum Singapore (2002)
6. Lloyd, S., Fillingham, D., Lampard, R., Orlowski, S., Weigelt, J.: CA-CA Interoperability, White Paper (March 2001)
7. Guo, Z., Okuyama, T., Marion Jr., R.F.: A New Trust Model for PKI Interoperability. In: Proceedings of the Joint International Conference on Autonomic and Autonomous Systems and International Conference on Networking and Services, ICAS/ICNS 2005 (2005)
8. Finin, T., Fritzson, R., McKay, D., McEntire, R.: KQML as an agent communication language. In: Proceedings of the Third International Conference on Information and Knowledge Management, CIKM 1994, pp. 456–463 (1994)
9. He, Q., Sycara, K.P., Finin, T.W.: Personal Security Agent: KQML-based PKI. In: Proceedings of the Second International Conference on Autonomous Agents, AGENTS 1998, pp. 377–384 (1998)
10. Hu, Y.J., Tang, C.W.: Agent-Oriented Public Key Infrastructure for Multi-Agent E-service. In: Palade, V., Howlett, R.J., Jain, L. C. (eds.) KES 2003. LNCS, vol. 2773, pp. 1215–1221. Springer, Heidelberg (2003)
11. Hu, Y.J.: Trusted Agent-Mediated E-Commerce Transaction Services via Digital Certificate Management. Electronic Commerce Research 3, 221–243 (2003)
12. Fugkeaw, S., Manpanpanich, P., Juntapremjitt, S.: Multi-Application Authentication based on Multi-Agent System. IAENG International Journal of Computer Science J 33(2), 1316–1321 (2007)
13. Fugkeaw, S., Manpanpanich, P., Juntapremjitt, S.: A Robust Sign-On Model based on Multi-Agent System and PKI. In: Proceedings of the Sixth International Conference on Networking, ICN 2007 (2007)
14. Randy, C.M., Joseph, G.T.: E-Commerce Security Issues. In: Proceedings of the 35th Hawaii International Conference on System Sciences (2002)
15. Zhang, Z., Zhang, C.: Basics of Agents and Multi-agent Systems. In: Zhang, Z., Zhang, C. (eds.) Agent-Based Hybrid Intelligent Systems. LNCS (LNAI), vol. 2938, pp. 29–39. Springer, Heidelberg (2004)
16. Yannis, L., Finin, T., Yun, P.: Agent Communication Languages: The Current Land-scape. IEEE Intelligent System 14(2), 45–52 (1999)
17. Foundation for Intelligent Physical Agents: FIPA specifications, http://www.fipa.org/ (accessed October 31, 2012)
18. Shimaoka, M., Hastings, N., Nielsen, R.: Network Working Group Request for Comments: 5217 Category: Informational, http://www.ietf.org/rfc/rfc5217.txt (accessed November 1, 2012)
19. Perlman, R.: An Overview of PKI Trust Models. IEEE Network, 38–43 (November/December 1999)

Using Bees Hill Flux Balance Analysis (BHFBA) for *in silico* Microbial Strain Optimization

Yee Wen Choon[1], Mohd Saberi Bin Mohamad[1], Safaai Deris[1],
Rosli Md. Illias[2], Lian En Chai[1], and Chuii Khim Chong[1]

[1] Artificial Intelligence and Bioinformatics Research Group,
Faculty of Computer Science and Information Systems,
Universiti Teknologi Malaysia,
Skudai, 81310 Johor, Malaysia
ywchoon2@live.utm.my,
{saberi,safaai}@utm.my,
{lechai2,ckchong2}@live.utm.my
[2] Department of Bioprocess Engineering,
Faculty of Chemical Engineering,
Universiti Teknologi Malaysia, Skudai,
81310 Johor, Malaysia
r-rosli@utm.my

Abstract. Microbial strains can be manipulated to improve product yield and improve growth characteristics. Optimization algorithms are developed to identify the effects of gene knockout on the results. However, this process is often faced the problem of being trapped in local minima and slow convergence due to repetitive iterations of algorithm. In this paper, we proposed Bees Hill Flux Balance Analysis (BHFBA) which is a hybrid of Bees Algorithm, Hill Climbing Algorithm and Flux Balance Analysis to solve the problems and improve the performance in predicting optimal sets of gene deletion for maximizing the growth rate and production yield of desired metabolite. *Escherichia coli* is the model organism in this paper. The list of knockout genes, growth rate and production yield after the deletion are the results from the experiments. BHFBA performed better in term of computational time, stability and production yield.

Keywords: Bees Algorithm, Hill Climbing, Flux Balance Analysis, Microbial Strains, Optimization.

1 Introduction

Microbial strains optimization has become popular in genome-scale metabolic networks reconstructions recently as microbial strains can be manipulated to improve product yield on desired metabolites and also improve growth characteristics [1]. Reconstructions of the metabolic networks are found to be very useful in health, environmental and energy issues [2]. The development of computational models for

A. Selamat et al. (Eds.): ACIIDS 2013, Part I, LNAI 7802, pp. 375–384, 2013.

simulating the actual processes inside the cell is growing rapidly due to vast numbers of high-throughput experimental data.

Many algorithms were developed in order to identify the gene knockout strategies for obtaining improved phenotypes. The first rational modeling framework (named OptKnock) for introducing gene knockout leading to the overproduction of a desired metabolite was developed by Burgard *et al.*, 2003 [3]. OptKnock identifies a set of gene (reaction) deletions to maximize the flux of a desired metabolite with the internal flux distribution is still operating such that growth is optimized.

OptKnock is implemented by using mixed integer linear programming (MILP) to formulate a bi-level linear optimization that is very promising to find the global optimal solution. OptGene is an extended approach of OptKnock which formulates the *in silico* design problem by using Genetic Algorithm (GA) [4]. Meta-heuristic methods are capable in producing near-optimal solutions with reasonable computation time, furthermore the objective function that can be optimized is flexible. SA is then implemented to allow the automatic finding of the best number of gene deletions for achieving a given productivity goal [5]. However, the results are not yet satisfactory.

A hybrid of BA and FBA was proposed by Choon *et al.*, 2012 [6], it showed a better performance in predicting optimal gene knockout strategies in term of growth rate and production yield. Pham *et al.*, 2006 [7] introduced Bees Algorithm (BA), is a typical meta-heuristic optimization approach which has been applied to various problems, such as controller formation [8], image analysis [9], and job multi-objective optimization [10]. BA is based on the intelligent behaviours of honeybees. It locates the most promising solutions, and selectively explores their neighbourhoods looking for the global maximum of the objective function. BA is efficient in solving optimization problems according to the previous studies [7, 10]. However, due to the dependency of BA on random search, it is relatively weak in local search activities [11]. Hence, BHFBA is proposed to improve the performance of BAFBA as Hill climbing algorithm is a promising algorithm in finding local optimum. This paper shows that BHFBA is not only capable in solving larger size problems in shorter computational time but also improves the performance in predicting optimal gene knockout strategy than previous works. In this work, we present the results obtained by BHFBA in two case studies where *Escherichia Coli (E.coli)* iJR904 model is the target microorganisms [12]. In addition, we also conduct a benchmarking to test performance of the hybrid of Bee algorithm and Hill climbing algorithm.

2 Bees-Hill Flux Balance Analysis (BHFBA)

In this paper, we proposed BHFBA in which BAFBA is only applied to identify optimal gene knockout strategies recently. Fig. 1 shows the flow of BAFBA while Fig. 2 shows our proposed BHFBA. The important steps are explained in the following subsections.

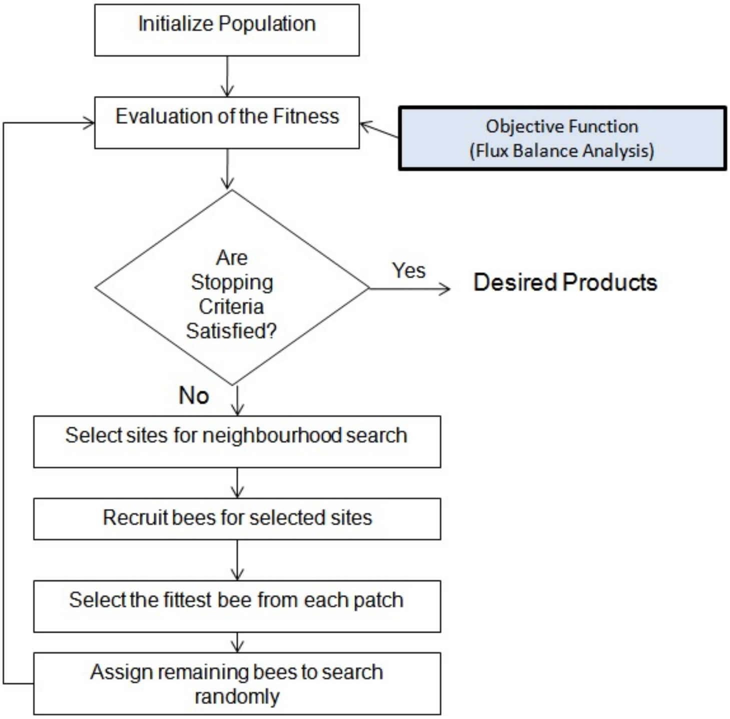

Initialize Population
Evaluation of the Fitness
Objective Function (Flux Balance Analysis)
Are Stopping Criteria Satisfied?
Yes
Desired Products
No
Select sites for neighbourhood search
Recruit bees for selected sites
Select the fittest bee from each patch
Assign remaining bees to search randomly

Fig. 1. BAFBA Flowchart

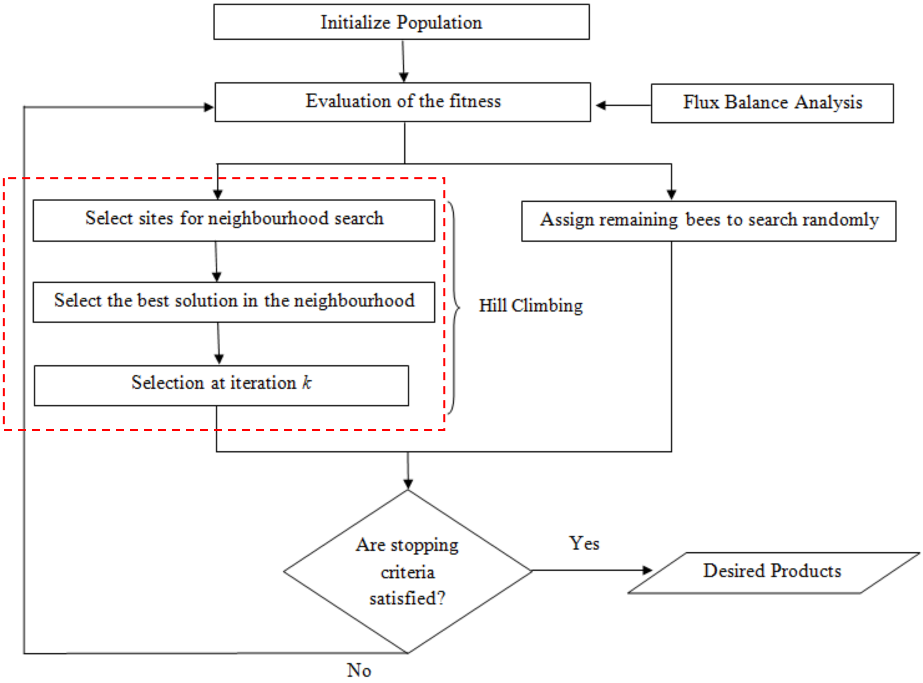
2.1 Model Pre-processing

The model is pre-processed through several steps based on biology assumptions as well as computational approaches to reduce the search space as while as increase the accuracy. Lethal reactions such as the genes that are found to be lethal *in vivo*, but not *in silico*, should be removed to improve the quality of the results. The results are invalid if a lethal reaction is deleted. The following are the details of computational pre-processing steps to the model [5].

a. Fluxes that are not associated with any genes, such as the fluxes related to external metabolites and exchange fluxes that represent transport reaction should not be involved in the process. These fluxes do not have a biological meaning thus they should not be knocked out.
b. Essential genes that cannot be deleted from the microorganism's genome need to be removed. The search space for optimization is reduced due to that these genes should not be considered as targets for deletion. A linear programming problem is defined by setting the corresponding flux to 0, while maximizing the biomass flux for each gene in the microorganism's genome. If the biomass flux result from the Linear Programming algorithm is zero (or near zero) then the gene is marked as essential. This biological meaning of this fact is that the microorganism is unable to

survive without this gene. This process does not suggest any changes to the model like the previous one, but provides favorable information for the optimization algorithms. With the help of biologists, the list of essential genes can be manually edited to include genes that are known to be essential *in vivo*, but not *in silico*.

c. Given the constraints of the linear programming problem, the fluxes need to be removed if the fluxes cannot exhibit values different from 0. Two linear programming are solved for every reaction in the model: the first is to define the flux over that reaction as the maximization target, while the second is to set the same variable as minimization target. If the objective function is 0 for both problems, then the variable is removed from the model.

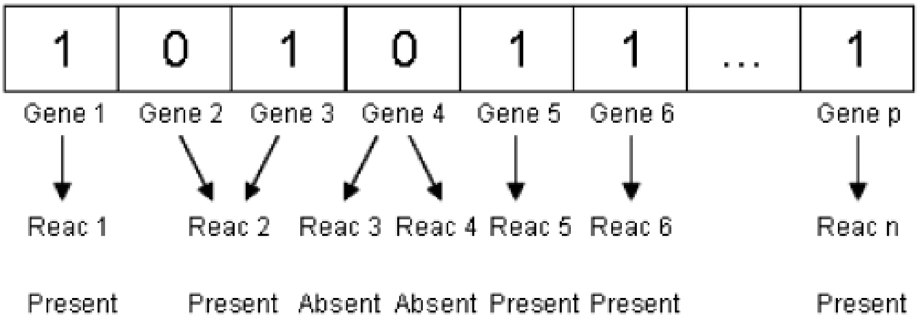

Initialize Population
Evaluation of the fitness
Flux Balance Analysis
Select sites for neighbourhood search
Assign remaining bees to search randomly
Select the best solution in the neighbourhood
Hill Climbing
Selection at iteration *k*
Are stopping criteria satisfied?
Yes
Desired Products
No

Note: Red-dotted box is Hill Climbing algorithm.

Fig. 2. BHFBA Flowchart

2.2 Bee Representation of Metabolic Genotype

One or more genes can be discovered in each reaction in a metabolic model. In this paper, each of those genes is represented by a binary variable indicating its absence or presence (0 or 1), these variables form a 'bee' representing a specific mutant that lacks some metabolic reactions when compared with the wild type (Fig. 3.)



Note: Reac represents reaction.

Fig. 3. Bee representation of metabolic genotype

2.3 Initialization of the Population

The algorithm starts with an initial population of n scout bees. Each bee is initialized as follows: assume that a reaction with n genes. Bees in the population are initialized by setting present or absent status to each gene randomly. Initialization of the population is done randomly so that all bees in the population have an equal chance of being selected. The result might not truly reflect the population if it is done with bias setting.

2.4 Scoring Fitness of Individuals

Each site is given a fitness score that determines whether to recruit more bees or should be abandoned. In this work, we used FBA to calculate the fitness score for each site and the equation is as follow:

Maximize Z, where

$$Z = \sum c_i v_i = \mathbf{c.v} \tag{1}$$

where c = a vector that defines the weights for of each flux.

Cellular growth is defined as the objective function Z, vector *c* is used to select a linear combination of metabolic fluxes to include in the objective function, *v* is the flux map and *i* is the index variable (*1, 2, 3, ..., n*). After optimizing the cellular growth, mutant with growth rate more than 0.1 continues the process by minimizing and maximizing the desired product flux at fixed optimal cellular growth value. Hence, we can enhance yield of our desired products at fixed optimal cellular growth. Production yield is the maximum amount of product that can be generated per unit of substrate. The following shows the calculation for production yield:

$$\text{Production yield} = (\text{production rate}_{\text{production}})/(\text{consumption rate}_{\text{substrate}})$$
$$(\text{mmol/mmol})(\text{gm/gm}) \tag{2}$$

where mmol = millimole and gm is gram.

We used Biomass-product coupled yield (BPCY) as the fitness score in this work, the calculation for BPCY is as follow:

$$\text{BPCY} = \text{product yield} * \text{growth rate } (\text{mmol}(\text{mmol}*\text{hr})^{-1})(\text{gm } (\text{gm} * \text{hr})^{-1}) \quad (3)$$

where mmol is millimole, hr is hour and gm is gram.

2.5 Neighbourhood Search (Hill Climbing Algorithm)

The algorithm carries out neighbourhood searches in the favored sites (*m*) by using Hill climbing algorithm. Hill climbing is an iterative algorithm that starts with an arbitrary solution to a problem, then attempts to find a better solution by incrementally changing a single element of the solution. In this paper, the initial solution is the *m* favored sites from the population initialized by BA. The algorithm starts with the solution and makes small improvements to it by adding or reducing a bee to the sites. User defined the value of initial size of patches (ngh) and uses the value to update site (*m*) which is declared in the previous step to search in neighbourhood area. In this paper, *m* is equal to 15 and ngh is equal to 30, the values are obtained by conducting a small number of trials with the range of 10 to 25 and 20 to 35 respectively. This step is important as there might be better solutions than the original solution in the neighbourhood area.

2.6 Randomly Assigned and Termination

The remaining bees in the population are sent randomly around the search space to scout for new feasible solutions. This step is done randomly to avoid overlooking the potential results that are not in the range. These steps are repeated until either the maximum loop value is met or the fitness function has converged. At the end, the colony generates two parts to its new population – representatives from each selected patch and other scout bees assigned to perform random searches.

3 Results and Discussion

In this paper, *E. coli i*JR904 is used to test on the operation of BAFBA [12]. The model contains 904 genes, 931 unique biochemical reactions, and 761 metabolites. The model is pre-processed through several steps based on biology assumptions and computational approaches before BHFBA is applied. This resulted in the size of the model is reduced to 667 reactions. Lethal reactions such as the genes that are found to be lethal *in vivo*, but not *in silico*, are not included as the possible targets in BHFBA. The reason to remove lethal reaction is that the microorganism is unable to survive without this reaction. The *E.coli* simulations are performed for aerobic minimal media conditions. The glucose uptake rate are fixed to 10 mmol/gDW/hr while a set non-growth associated maintenance of 7.6 mmol ATP/gDW/hr. The experiments are carried out by using a 2.3 GHz Intel Core i7 processor and 8 GB DDR3 RAM computer.

Table 1 and Table 2 summarize the results obtained from BHFBA for succinic acid and lactic acid production from *E.coli*. As shown from the results, this method has produced better results to the previous works in term of growth rate and BPCY meanwhile potential genes which can be removed are identified [3][4][5].

Table 1. Comparison between different methods for production of Succinic acid in *E.coli*

Method	Growth Rate (1/hr)	BPCY	List of knockout genes
BHFBA	0.7988	0.93656	PTAr**, RPE, SUCD1i
BAFBA [10]	0.62404	0.66306	FUM, PTAr**, TPI**
SA + FBA [5]	N/A	0.39850	ACLD19*, DRPA, GLYCDx, F6PA, TPI**, LDH_D2, EDA, TKT2, LDH_D-
OptKnock [3]	0.28	N/A	ACKr, PTAr**, ACALD*

Note: The shaded column represents the best result. N/A – Not Applicable. * Common genes for all methods. ** Common genes in either 2 methods. BPCY is in gram (gram-glucose.hour)$^{-1}$.

Table 1 shows that BHFBA performed better than the previous works with growth rate 0.7988 and BPCY 0.93656. Knocking out succinate dehydrogenase (SUCD1i) interrupts the formation from succinic acid to fumarate. Without the conversion from succinic acid to fumarate, production yield of succinic is improved. Next, phosphotransacetylase (PTAr) is removed, according to Burgard *et al.*, 2003[3], the mutants can grow anaerobically on glucose by producing lactate. In the next step, ribulose-5-phosphate-3-epimerase (RPE) is suggested to knockout. This knockout involves the inflow reaction of ammonium. As stated in Bohl *et al.*, 2010 [13], the utilization of nitrate as electron acceptor and ammonium source under anaerobic conditions can improve succinate production.

Table 2 shows the results of BHFBA and previous works. BHFBA obtained a better BPCY that is 5.2241 than the previous work. However, the growth rate is slightly lower than BAFBA due to BHFBA finds a solution with higher BPCY with the condition that the growth rate is more than 0.1. The deletion of fructose bisphosphatase and phosphoglycerate kinase decreased the efficiency of gluconeogenesis which resulted in the increment concentration of phosphoenolpyruvate. Phosphoenolpyruvate is then converted into pyruvate and continues to convert into lactic acid. Knocking out acetaldehyde dehydrogenase which catalyze the conversion of acetaldehyde into acetic acid eliminated the competiting product, acetic acid. In consequence the yield of lactic acid is improved.

Table 3 shows the computational time comparison between BHFBA and BAFBA for 1000 iterations. The average computational time of BHFBA improved 69% of the BAFBA result for 1000 iterations.

Table 2. Comparison between different methods for production of Lactic acid in *E.coli*

Method	Growth Rate (1/hr)	BPCY	List of knockout genes
BHFBA	0.62501	5.2241	FBP, PGK, ACALD*
BAFBA [10]	0.86381	2.1677	ACALD*, ACKr**, GLUN, GND, ME1
SA + FBA [5]	N/A	0.39850	ACLD19*, DRPA, GLYCDx, F6PA, TPI, LDH_D2, EDA, TKT2, LDH_D-
OptKnock [3]	0.28	N/A	ACKr**, PTAr, ACALD*

Note: The shaded column represents the best result. N/A – Not Applicable. * Common genes for all methods. ** Common genes in either 2 methods. BPCY is in gram (gram-glucose.hour)$^{-1}$.

Table 3. Comparison between average computational time of BHFBA and BAFBA for 1000 iterations

Method	Computation Time (s)
BHFBA	3223
BAFBA	10253

In addition, since BA and Hill Climbing algorithm is a new hybrid algorithm. Hence, we conducted a benchmarking to test performance of a hybrid of BA and Hill Climbing algorithm (BH). As BA is looking for the maximum, the functions are inverted before the algorithm is applied. The De Jong and Martin & Gaddy functions are used in this paper. Table 4 shows the mathematical representation of the functions. Table 5 shows mean and standard deviation (STD) of the both functions, De Jong and Martin & Gaddy, tested on both original BA and BH. As seen from the results, both BHFBA and BH performed better than other algorithms. It can be concluded that the capability of Hill Climbing algorithm in finding local optimum improved the performance of the original BA. The original BA with the problem of repetitive iterations of the algorithm in local search where each bee keep searching until the best possible answer is reached. Our proposed BHFBA solved the problem by implementing Hill Climbing algorithm into the local search part. Hill Climbing algorithm is a powerful local search algorithm which attempts to find a better solution by incrementally changing a single element of the solution until no further improvements can be found, the search process is recorded so the process is not repeated. Furthermore, one of the advantages of Hill Climbing algorithm is it can return a valid solution even if it is interrupted at any time before it ends.

Table 4. Mathematical representation of De Jong and Beale functions

Name	Mathematical representation
De Jong	$\max F = (3905.93) - 100(x_1^2 - x_2)^2 - (1 - x_1)^2$
Martin & Gaddy	$\min F = (x_1 - x_2)^2 + ((x_1 + x_2 - 10) / 3)^2$

Table 5. Obtained fitness value of both De Jong and Beale functions

Function	Mean		STD	
	BA	BH	BA	BH
De Jong	3.91e+03	3.90e+03	0.000504	4.79e-13
Martin & Gaddy	11.1083	11.1111	0.002797	0

4 Conclusion and Future Works

In this paper, BHFBA is proposed to predict optimal sets of gene deletion to maximize the production of desired metabolite. Experimental results on *E. coli i*JR904 model obtained from literature showed that BHFBA is effective in generating optimal solutions to the gene knockout prediction, and is therefore a useful tool in Metabolic Engineering [12]. We are interested in applying other fitness functions in BHFBA such as minimization of metabolic adjustment (MOMA) and regulatory on/off minimization (ROOM) to further improve the performance of BHFBA. Besides that, BA employs many tunable parameters which are difficult for the users to determine so it is important to find ways to help the users choose suitable parameters.

Acknowledgement. Institutional Scholarship MyPhD provided by the Ministry of Higher Education of Malaysia finances this work. We also would like to thank Universiti Teknologi Malaysia for supporting this research by the UTM GUP research grants (Vot numbers: QJ130000.7107.01H29 and QJ130000.7123.00H67).

References

1. Feist, A.M., Herrgård, M.J., Thiele, I., Reed, J.L., Palsson, B.Ø.: Reconstruction of biochemical networks in microorganisms. Nat. Rev. Microbiol. 72, 129–143 (2009)
2. Chandran, D., Copeland, W.B., Sleight, S.C., Sauro, H.M.: Mathematical modeling and synthetic biology. Drug Discovery Today Disease Models 5(4), 299–309 (2008)
3. Burgard, A.P., Pharkya, P., Maranas, C.D.: OptKnock: A bilevel programming framework for identifying gene knockout strategies for microbial strains optimization. Biotechnol. Bioeng. 84, 647–657 (2003)
4. Patil, K.R., Rocha, I., Förster, J., Nielsen, J.: Evolutionary programming as a platform for in silico metabolic engineering. BMC Bioinformatics 6, 308 (2005)
5. Rocha, M., Maia, P., Mendes, R., Pinto, J.P., Ferreira, E.C., Nielsen, J., Patil, K.R., Rocha, I.: Natural computation meta-heuristics for the in silico optimization of microbial strains. BMC Bioinformatics 9, 499 (2008)
6. Pham, D.T., Ghanbarzadeh, A., Koç, E., Otri, S., Zaidi, M.: The bees algorithm – a novel tool for complex optimization problems. In: Proceedings of the Second International Virtual Conference on Intelligent Production Machines and Systems, July 3-14 (2006)
7. Choon, Y.W., Mohamad, M.S., Deris, S., Chong, C.K., Chai, L.E., Ibrahim, Z., Omatu, S.: Identifying Gene Knockout Strategies Using a Hybrid of Bees Algorithm and Flux Balance Analysis For in silico Optimization of Microbial Strains. In: The 9th International Symposium on Distributed Computing and Artificial Intelligence, DCAI 2012. University of Salamanca, Spain (2012)

8. Pham, D.T., Darwish, A.H., Eldukhri, E.E.: Optimisation of a fuzzy logic controller using the bees algorithm. International Journal of Computer Aided Engineering and Technology 1(2), 250–264 (2006)
9. Olague, G., Puente, C.: The Honeybee Search Algorithm for Three-Dimensional Reconstruction. In: Rothlauf, F., Branke, J., Cagnoni, S., Costa, E., Cotta, C., Drechsler, R., Lutton, E., Machado, P., Moore, J.H., Romero, J., Smith, G.D., Squillero, G., Takagi, H. (eds.) EvoWorkshops 2006. LNCS, vol. 3907, pp. 427–437. Springer, Heidelberg (2006)
10. Pham, D.T., Ghanbarzadeh, A.: Multi-objective optimisation using the bees algorithm. Paper Proceedings of the Third International Virtual Conference on Intelligent Production Machines and Systems, July 2-13 (2007)
11. Cheng, M.Y., Lien, L.C.: A Hybrid Swarm Intelligence Based Particle Bee Algorithm For Benchmark Functions And Construction Site Layout Optimization. In: Proceedings of the 28th ISARC, Seoul, pp. 898–904 (2011)
12. Reed, J.L., Vo, T.D., Schilling, C.H., Palsson, B.O.: An expanded genome-scale model of Escherichia coli K-12 (iJR904 GSM/GPR). Genome Biology 4, 54 (2003)
13. Bohl, K., Figueiredo, L.F.: d., Hadicke O., Klamt S., Kost C., Schuster S., Kaleta C: CASOP GS: Computing intervention strategies targeted at production improvement in genome-scale metabolic networks. In: The 5th German Conference on Bioinformatics 2010, September 20-22 (2010)

Multiple Gene Sets for Cancer Classification Using Gene Range Selection Based on Random Forest

Kohbalan Moorthy[*], Mohd Saberi Bin Mohamad, and Safaai Deris

Artificial Intelligence & Bioinformatics Research Group, Faculty of Computer Science and Information Systems, Universiti Teknologi Malaysia, 81310 Skudai, Johor, Malaysia
kohbalan@gmail.com, {saberi,safaai}@utm.my

Abstract. The advancement of microarray technology allows obtaining genetic information from cancer patients, as computational data and cancer classification through computation software, has become possible. Through gene selection, we can identify certain numbers of informative genes that can be grouped into a smaller sets or subset of genes; which are informative genes taken from the initial data for the purpose of classification. In most available methods, the amount of genes selected in gene subsets are dependent on the gene selection technique used and cannot be fine-tuned to suit the requirement for particular number of genes. Hence, a proposed technique known as gene range selection based on a random forest method allows selective subset for better classification of cancer datasets. Our results indicate that various gene sets assist in increasing the overall classification accuracy of the cancer related datasets, as the amount of genes can be further scrutinized to create the best subset of genes. Moreover, it can assist the gene-filtering technique for further analysis of the microarray data in gene network analysis, gene-gene interaction analysis and many other related fields.

Keywords: Gene Selection, Cancer Classification, Random Forest, Gene Expression, Microarray Data.

1 Introduction

Vast amount of data generation have led to the development of many sophisticated methods and tools for visualization and analysis of data. These huge amounts of data, particularly for the biological analysis and interpretation, are made available through microarray technology [1]. Microarray technology allows continuous analysis and interpretation of the expression levels present in the observed genes from microarray data. Analyzing microarray data is a challenging task, as the high dimensionality of the data requires large processing power with sufficient amount of memory resources. Furthermore, microarray technology allows the expansion of information of the sample itself, where detailed insights of the data can be used for gene regulation and identification based on gene expression data [2]. In addition, it has been used in studies

[*] Corresponding author.

A. Selamat et al. (Eds.): ACIIDS 2013, Part I, LNAI 7802, pp. 385–393, 2013.

related to cancer classification, identification of relevant genes for diagnosis or therapy and investigation of drug effects on cancer prognosis [3].

Biologists require accurate predictive tools as well as group of relevant genes for biomarkers in cancer identification [4]. Cancer informatics has been expected to be a part of the advancement in the identification and validation of biomarkers through the combine interdisciplinary fields, which expands from the bioinformatics [5]. Prior to classification, performing gene selection allows grouping of relevant genes into a subset. Some of the main reasons for performing gene selection are to avoid over fitting for improved model performance, to gain faster and less costly models and lastly to dig deeper into the data generation processes [6].

Gene selection approach is divided into three main categories, which are filter based approach, wrapper based approach and embedded based approach [7]. Filter based approach is defined as when the gene selection process is carried out independently of the classification algorithms. If the classifier is being used to evaluate every selected subset of the gene selection process throughout the entire classification process, then it is known as a wrapper based approach [8]. Embedded approach uses the same classifier dependent selection as the wrapper based approach, except that it has better computational complexity. According to Wong, Leckie and Kowalczyk [9], filter based approach performs gene selection without any dependence on the classifier being chosen, which may not be sufficient enough to generate higher accuracy in classification as those of wrapper and embedded approaches, which have certain degree of dependencies with the classifier algorithm being used. In spite of that, wrapper based approach is not preferred in sample classification due to huge combination of genes subset required to be examined. Moreover, the wrapper method requires high computation time and it is much slower in determining the best subset of genes [10].

Accurately categorizing the selected genes into their respective class as into normal or tumor is known as the process of binary classification. Classifier can be defined as an artificial intelligence device, which has the potential to make classification [11]. In usual classification scenario, most developed algorithms focus on maximizing the overall correct predictors in order to gain higher accuracy in classification even though there is an imbalance in the different class size [12]. Some examples of classifiers are support vector machines (SVM), neural network (NN), k-nearest neighbor (kNN) and classification tree.

In genetic associated studies, Random Forest has been used widely for both classification and gene selection [13]. Random Forest was first developed by Breiman [14] for the purpose of classification, regression, clustering and also survival analysis. In this field, the practice and application of gene ranking are according to the genes contribution towards a disease. Random forest has been one of the favored methods used in gene importance measurement for gene ranking and selection. Diaz-Uriarte and Alvarez de Andres [15] had proposed a gene selection and classification based on Random Forest for the first time as an embedded approach. Besides that, Random Forest algorithm is effective in predicting samples, as well as revealing interactions among the genes. Additionally, a limiting value is achieved as the number of trees set in the Random Forest is increased continuously, making it an ideal error predictor with no over fitting occurrence of the data. In Random Forest, trees are grown, and from the training sample, each tree grows without pruning from the actual data based on random gene selection.

For the creation of gene expression profiles, many researchers are continuously seeking for state of the art classification algorithms that can provide better accuracy. Gene selection has played a vital role in increasing the classification accuracy for cancer related disease but most of the gene selection techniques available are unrelated to the classification algorithm. Moreover, the amount of genes selected in gene sub-sets are dependent on the gene selection technique used and cannot be fine-tuned to suit the requirement for particular number of genes. Hence, we propose a technique on gene range selection based on a Random Forest method for selective subset, leading to better classification of cancer datasets.

In this article, we begin by describing the methodology section where the proposed technique is briefly explained; followed by the result and discussion section, where the main characteristics of the datasets are explained, and the complete analysis of the findings is presented. Comparisons with previous similar research papers are also presented to further justify the improvement achieved using the proposed technique. Lastly, the future works and conclusion of this article are presented.

2 Methodology

Diaz-Uriarte and Alvarez de Andres [15] first proposed the gene selection through Random Forest algorithm. Moorthy and Mohamad [16] then proposed an improved version of the gene selection. In this research, we propose an improvement on the existing gene selection technique based on the Random Forest method, which is gene range selection. Most existing techniques and methods used for gene selection do not reveal the amount of genes selected for training the classifier. Moreover, the selected subset of genes is very dependent on the gene selection technique and does not have the capability to tune and finalize the amount of the selected genes for extended usage in other related fields, such as gene network analysis, gene-gene interaction analysis, and gene annotations. Besides that, most of the gene selection techniques produce constant output of genes for the use of the classification algorithms. Therefore, there are no possibilities of tweaking that particular gene selection technique to evaluate the different output performance of the classifier.

Through this research, an enhancement to the gene selection technique is introduced to provide the flexibility and options to generate different gene sets with better accuracy, as well as the ability to control the amount of genes required on each gene subset. The idea of this improvement focuses on allowing the gene selection algorithm to test and evaluate a certain range of genes from the overall dataset and evaluate the final classification accuracy. Furthermore, it allows analysis and comparison of different gene subsets towards the classification accuracy. The main reason for introducing this improved gene selection technique is to provide various gene range selections in any particular selected gene subset for better cancer classification. Moreover, it is also to allow other researchers to further tweak and select their desire range of genes in any particular gene subset which can provide better analysis capability in other research areas.

In order to achieve the proposed gene section technique, modification to the steps in the backward elimination process were carried out to accept inputs of selective range of genes, which were taken as minimum value (MinVar) and maximum value (MaxVar).

Prior to that, the cancer dataset were represented in two different forms of dataset information (Data) and to class the dataset to (Class). While performing the back-ward elimination process, a new subset was generated and evaluated where the previous error rates obtained (p.mean) were compared with the current error rates obtained (c.mean), and if there was a reduction, then the previous best would be replaced with the current best. Once the best subset of genes was determined and the required number of genes was satisfied, we then used the gene subset (bestSub) for the classification process. A complete flow of the gene range technique is been presented in Figure 1, where the dotted line represents the changes made to achieve the range selection.

```
Gene Range Technique
Input: Data, Class, MinVar and MaxVar
Output: Selected genes and error rates
while backward elimination process = true do
   removes fraction of genes;
   test and evaluate remaining genes;
   c.mean = current error rates;
   p.mean = previous error rates;
   if c.mean <= p.mean
       p.mean = c.mean;
       selVar = current subset of genes;
       if selVar <= MaxVar and selVar >= MinVar
           bestSub = selVar;
       end if
   end if
   if selVar < MinVar
       break;
   end if
end while
```

Fig. 1. Pseudo code for the gene range selection developed for controlled amount of selected genes in a particular subset

3 Results and Discussion

In this research, we used cancer related datasets, which were gene expression dataset obtained through the microarray technology. The datasets involved in this research could be grouped into various cancer types, which includes adenocarcinoma, breast cancer, colorectal cancer, leukemia and prostate cancer. These cancer datasets were primarily binary, which are known as two-class dataset and consists of both the normal, and tumor based patient samples.

The cancer datasets used for this research were in text file format and had been pre-formatted to suit the software. For each of the cancer dataset, they have two main text files, which were class file and data file. The class file contained the information to identify the data file according to normal or tumor samples. The data file consists of numerical values, where the rows represent the total number of genes in any particular cancer dataset and the columns represent the total number of patients. The detailed description of the cancer dataset is presented in the Table 1, where the number of genes, patients and the main reference of the data are listed.

Table 1. Main characteristics of the cancer dataset used in this research

Dataset Name	Genes	Patients	Reference
Adenocarcinoma	9868	76	[17]
Breast	4869	77	[18]
Colon	2000	62	[19]
Leukemia	3051	38	[20]
Prostate	6033	102	[21]

The complete analysis for the selected cancer datasets had been tabulated according to selected gene range settings, and both the number of genes in a subset and error rates were obtained. The selected gene range had been set to into four different partitions as to 2 to 10 genes for the first range, 10 to 50 genes for the second range, 50 to 250 genes for the third range and the final range from 250 genes to the maximum number of genes present in any particular dataset.

The minimum of two genes were preset, as each gene from the tumor and normal was required as the minimum informative genes for the classification purpose. The selected gene range settings executed were used to determine the local optimum genes subset for the entire dataset and each subset could be selected to be further used into the classification process. In terms of the error rates calculation, the .632+ Bootstrap error rates from Efron and Tibshirani [22] had been applied. The complete result is presented in Table 2.

From the results obtained, we can see that the best number of genes for Adenocarcinoma dataset was 12 genes, with the classification error rates obtained as low as 0.1801 compared to other selected range. Even though with 222 genes, the error rates obtained were the lowest among all the selected range, which was 0.1745, the number of genes in this particular subset was huge and did not compensate the 3% improvement in accuracy compared to the ratio of the genes. This could be the indication that there were some genes in the subset of 222 genes, which might be useful in increasing the overall accuracy. Similar case happened to the Breast cancer dataset as we can see that the lowest error rates obtained were 0.3249 with 214 genes in the selected subset. The recommended subset of genes for this dataset would be 56 genes as the error rate obtained was 0.3257, which was similar but slightly higher than the best error rate and the difference was only 0.2%, yet the difference in the number of genes is 214 genes over 56 genes.

Apart from that, Colon cancer dataset and Leukemia dataset showed a similar gene range selection, as the best gene subset for Colon cancer dataset consisted of 18 genes whereas 9 genes were obtained for the Leukemia dataset. The lowest error rates obtained for Colon cancer dataset and Leukemia dataset were 0.1539 and 0.0753, respectively. Most probably, both this datasets had much lesser informative genes in overall compared to other datasets. Therefore, higher number of genes would only affect the classification accuracy and increased the error rates. For Prostate cancer

Table 2. Classification error rates of the cancer dataset based on gene range selection technique where the shaded area represents lowest error rates

Gene Range	Adenocarcinoma		Breast		Colon		Leukemia		Prostate	
	*No of Genes	Error Rates	*No of Genes	Error Rates	*No of Genes	Error Rates	*No of Genes	Error Rates	*No of Genes	Error Rates
2 – 10	2	0.1871445	6	0.3287878	3	0.1655999	2	0.07934855	2	0.06249576
10 – 50	12	0.1801157	29	0.3311092	18	0.1539477	9	0.07529078	18	0.06060309
50 – 250	58	0.1844413	56	0.3256622	56	0.1681416	44	0.08684319	109	0.06014227
250 – max**	222	0.1744816	214	0.3249034	214	0.1625876	210	0.08636869	212	0.05849378

* Total genes present in any particular selected subset.
** All genes in the dataset.

dataset, the average error rate obtained was 0.06 and with the different gene range selection, there were no significant differences. Even though the best gene subset contained 212 genes, based on the error rates differences, the preferred gene subset would be with 18 genes. This could be due to the amount of neutral genes, which did not contribute enough in the classification. With the various selection ranges, the best subset from each range partition had been used for the random forest classifier to obtain the highest possible accuracy, which is presented in the Figure 2.

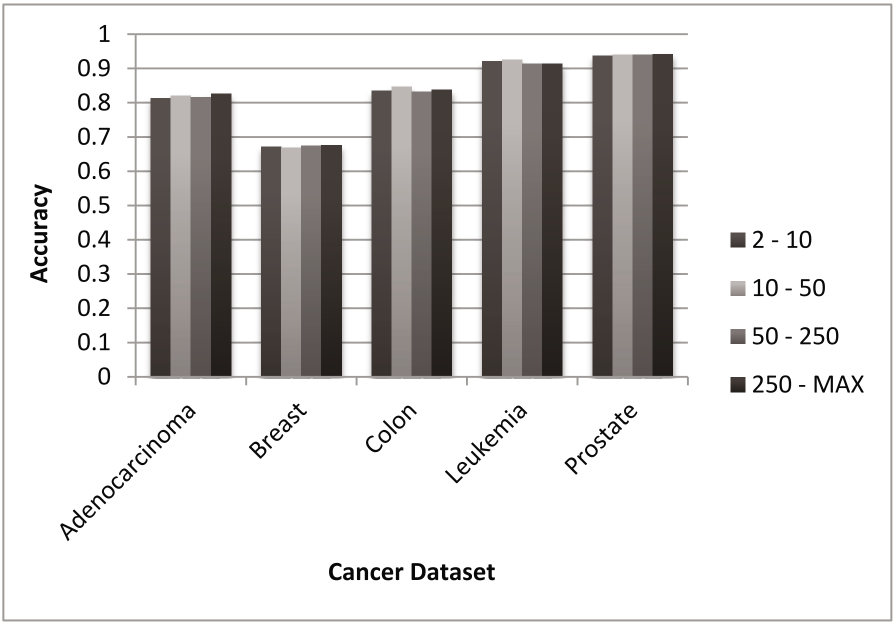


Fig. 2. Comparison of different gene range selection towards the overall classification accuracy of the cancer datasets

From our analysis, we could deduce that the suitable range for informative genes was at 10 – 50 genes range, as most of the dataset shown better or higher accuracy in this range. Even though the difference was not intermittent in terms of accuracy, but the amount of genes were either too less or too many for other selected ranges. However, other researchers may use the variance of the genes amount for subsequent analysis as well as a gene filtration for large datasets.

Besides that, the gene range selection can be altered to suit other requirements such as for the construction of gene network analysis, genes functional annotation through gene ontology and many more subsequent analyses.

4 Future Works

Cancer detection through Single nucleotide polymorphism (SNP) is a crucial stage in the prediction of cancer patients and it would be another step of advancement if the

Random Forest method can be altered to accept feeds from the SNP type microarray data in future. Besides that, the annotation of the selected genes and cross-referencing with genes databases could provide better understanding and validation of future predicted gene subsets.

5 Conclusion

The gene range selection technique has been tested with five different cancer datasets and the outcome of the classification has been presented in the result and discussion section. With the wide possibilities of gene subset selection, the accuracy of the classification based on the selected subsets has shown similar or better accuracy with no such fluctuation on the overall accuracy. This allows different range of genes to be selected from the entire datasets without deteriorating the classification accuracy.

Most gene selection techniques do not provide the actual number of genes in the selected subset, nor the flexibility to tune the amount of genes to be chosen in any particular gene subset prior to classification. We have shown a method of solution with the proposed gene range selection technique, which allows fine-tuning of the amount of genes selected in any particular gene subset without degrading the classification accuracy. Through the development of the gene range technique for the Random Forest gene selection, different subsets of genes with better classification accuracy have been listed for various use of gene expression analysis. The possibility for further analysis through gene network analysis, gene – gene interaction analysis and other related analysis is also made available, for the researchers may have their own preference of range of selection to obtain various sets of genes. This will not only allow controlling the amount of genes to be obtained but also provide accuracy of estimation based on the comparison of the selected genes.

Acknowledgements. Institutional Scholarship MyPhd provided by the Ministry of Higher Education of Malaysia finances this work. We also would like to thank Universiti Teknologi Malaysia for supporting this research by UTM GUP research grants (vot numbers: QJ130000.7123.00H67 and QJ130000.7107.01H29).

References

1. Paz, J.L., Seeberger, P.H.: Recent Advances and Future Challenges in Glycan Microarray Technology. In: Chevolot, Y. (ed.) Carbohydrate Microarrays, vol. 808, pp. 1–12. Humana Press (2012)
2. Pham, T.D., Wells, C., Crane, D.I.: Analysis of Microarray Gene Expression Data. Current Bioinformatics 1, 37–53 (2006)
3. Liew, A.W.-C., Law, N.-F., Yan, H.: Missing value imputation for gene expression data: computational techniques to recover missing data from available information. Briefings in Bioinformatics 12, 498–513 (2011)
4. Duval, B., Hao, J.-K.: Advances in metaheuristics for gene selection and classification of microarray data. Briefings in Bioinformatics 11, 127–141 (2010)
5. Wu, D., Rice, C., Wang, X.: Cancer bioinformatics: A new approach to systems clinical medicine. BMC Bioinformatics 13, 71 (2012)

6. Saeys, Y., Inza, I., Larrañaga, P.: A review of feature selection techniques in bioinformatics. Bioinformatics 23, 2507–2517 (2007)
7. Van Steen, K.: Travelling the world of gene–gene interactions. Briefings in Bioinformatics 13, 1–19 (2012)
8. Hua, J., Tembe, W.D., Dougherty, E.R.: Performance of feature-selection methods in the classification of high-dimension data. Pattern Recogn. 42, 409–424 (2009)
9. Wong, G., Leckie, C., Kowalczyk, A.: FSR: feature set reduction for scalable and accurate multi-class cancer subtype classification based on copy number. Bioinformatics 28, 151–159 (2012)
10. Nanni, L., Brahnam, S., Lumini, A.: Combining multiple approaches for gene microarray classification. Bioinformatics 28, 1151–1157 (2012)
11. Asyali, M.H., Colak, D., Demirkaya, O., Inan, M.S.: Gene Expression Profile Classification: A Review. Current Bioinformatics 1, 55–73 (2006)
12. Lin, W.-J., Chen, J.J.: Class-imbalanced classifiers for high-dimensional data. Briefings in Bioinformatics (2012)
13. Boulesteix, A.-L., Bender, A., Lorenzo Bermejo, J., Strobl, C.: Random forest Gini importance favours SNPs with large minor allele frequency: impact, sources and recommendations. Briefings in Bioinformatics 13, 292–304 (2012)
14. Breiman, L.: Random Forests. Mach. Learn. 45, 5–32 (2001)
15. Diaz-Uriarte, R., Alvarez de Andres, S.: Gene selection and classification of microarray data using random forest. BMC Bioinformatics 7, 3 (2006)
16. Moorthy, K., Mohamad, M.S.: Random forest for gene selection and microarray data classification. Bioinformation 7, 142–146 (2011)
17. Ramaswamy, S., Ross, K.N., Lander, E.S., Golub, T.R.: A molecular signature of metastasis in primary solid tumors. Nature Genetics 33, 49–54 (2003)
18. van 't Veer, L.J., Dai, H., van de Vijver, M.J., He, Y.D., Hart, A.A., Mao, M., Peterse, H.L., van der Kooy, K., Marton, M.J., Witteveen, A.T., Schreiber, G.J., Kerkhoven, R.M., Roberts, C., Linsley, P.S., Bernards, R., Friend, S.H.: Gene expression profiling predicts clinical outcome of breast cancer. Nature 415, 530–536 (2002)
19. Alon, U., Barkai, N., Notterman, D.A., Gish, K., Ybarra, S., Mack, D., Levine, A.J.: Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotide arrays. Proceedings of the National Academy of Sciences 96, 6745–6750 (1999)
20. Golub, T.R., Slonim, D.K., Tamayo, P., Huard, C., Gaasenbeek, M., Mesirov, J.P., Coller, H., Loh, M.L., Downing, J.R., Caligiuri, M.A., Bloomfield, C.D., Lander, E.S.: Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. Science 286, 531–537 (1999)
21. Singh, D., Febbo, P.G., Ross, K., Jackson, D.G., Manola, J., Ladd, C., Tamayo, P., Renshaw, A.A., D'Amico, A.V., Richie, J.P., Lander, E.S., Loda, M., Kantoff, P.W., Golub, T.R., Sellers, W.R.: Gene expression correlates of clinical prostate cancer behavior. Cancer Cell 1, 203–209 (2002)
22. Efron, B., Tibshirani, R.: Improvements on Cross-Validation: The .632+ Bootstrap Method. Journal of the American Statistical Association 92, 548–560 (1997)

Ubiquitous Positioning: Integrated GPS/Wireless LAN Positioning for Wheelchair Navigation System

Wan Mohd Yaakob Wan Bejuri[1], Wan Mohd Nasri Wan Muhamad Saidin[1],
Mohd Murtadha Bin Mohamad[1], Maimunah Sapri[2], and Kah Seng Lim[1]

[1] Faculty of Computing,
UniversitiTeknologi Malaysia
[2] Centre of Real Estate Studies,
UniversitiTeknologi Malaysia
{wanmohdyaakob,wanmohdnasri,lim0709}@gmail.com,
{murtadha,maimunahsapri}@utm.my

Abstract. The location determination in obstructed area can be very challenging especially when the Global Positioning System is blocked. Disable users will find it difficult to navigate directly on-site in such condition, particularly in obstructed environment. Sometimes, it needs to integrate with other sensors and positioning methods in order to determine the location with more intelligent, reliable and ubiquity. By using ubiquitous positioning, it provides the location technique inside the wheelchair navigation system that needed for disable people. In this paper, we utilizes the integration of wireless local area network and the Global Positioning System which receive signal strength from access point and at the same time, it retrieve Global Navigation System Satellite signal. This positioning information will be switched based on type of environment in order to ensure the ubiquity of wheelchair navigation system. Finally, we present our results by illustrating the performance of the system for an indoor/ outdoor environment set-up.

Keywords: Global Navigation System, Wireless LAN and Wheelchair Navigation.

1 Introduction

Knowledge of the location position is a common requirement for many people[1][2]. Over the last few years, many research and development has taken place concerning about location-based services (LBS), which could be supplemented and expanded with the help of ubiquitous methods, and be replaced in the future. The positioning and tracking of pedestrians in smart environment is achieved differently to the use of conventional navigation systems, as it is no longer only passive systems, which execute positioning on demand, that need to be considered. By finding the location positioning technologies on wheelchair navigation can help to navigate wheelchair in various environments[3][4]. The aim of wheelchair navigation is to locate and track a wheelchair accurately[5].

A. Selamat et al. (Eds.): ACIIDS 2013, Part I, LNAI 7802, pp. 394–403, 2013.

Nowadays, the usage of mobile device as wheelchair navigation devices has been widely known, since it is the coolest gadget that promote mobility, efficiency and effective to the end users [6][7]. Most of the high-end mobile device has been equipped with the Global Positioning System (GPS) navigation system that can generate advice for the wheelchair user to reach a desired destination. For example, if the user gets lost, the system can guide the user back on track. However, the navigation by using stand alone GPS was suffering in obstruction environment especially when the user is inside a building (indoor environment) [8] [9] [10][11]. Therefore, integration between GPS with other positioning method is needed, in order to determine positioning information with more pervasive, reliable and ubiquity. In addition, objects such as trees, high buildings, high walls and even people walking may constitute an obstruction to the signal. These obstructions sometimes fool the system into believing that the user has moved to another location; this usually happens in indoor environments, and makes it hard to estimate the user's position. Therefore, there is a need for an alternative method which ensures that users can locate themselves inside buildings as well as outdoors (for example, a visitor may want to find a friend in a complex office building). In this paper, we proposed the design of wheelchair navigation system in term on how to get positioning information in both inside building and also outdoor environment, by using integration of GPS and WLAN. The structure of the paper is as follows. Section 2 will present the reviews related research to GPS-based positioning and indoor-based positioning system. Section 3 will present an overview of our proposed method. The details of our result will be cover in section 4. Finally, conclusions are given in section 5.

2 Related Research

The main issue of most popular mobile navigation system; (GPS standalone), is its positioning accuracy. Furthermore, this technology makes the navigation system as overall so, it becomes difficult to be operated in obstructed environment; especially inside a building. Therefore, the concept of indoor positioning system has been introduced. The world first indoor WLAN positioning; known as RADAR [12], which adopts the nearest neighbor(s) in signal-space technique. The accuracy of the proposed system is around 2–3 m. Later, RADAR was enhanced by a Viterbi-like algorithm. The result is that the 50 percent of the RADAR system is around 2.37–2.65 m and its 90 percentile is around 5.93–5.97 m. Horus system [13], it is offered a joint clustering technique for location estimation, which uses the probabilistic method that been described previously. Each candidate location coordinate is considered as a class or category. The experiment results show that this technique can acquire accuracy of more than 90% to within 2.1 m. A grid-based Bayesian location-sensing system over a small region of their office building, achieving localization and tracking to within 1.5 m over 50% of the time. Nibble [14], one of the first system of this generation, used a probabilistic approach (based on Bayesian network) to estimate a device's location. In, Battiti et al. proposed a location determination method by using neural-network-based classifier by adopting multilayer perceptron (MLP) architecture and one-step secant (OSS) training method. They reported that only five samples of signal strengths in different locations are sufficient to get an average distance error of 3 m. By keeping the number of training examples increase will help decrease the average distance error to 1.5 m. At the same time, the research goes beyond by making positioning system more pervasive and ubiquity, the company known as Qualcomm, overcome the limitations of conventional GPS, and provide GPS indoors technique with an average of 5–50 m accuracy in most indoor environments. A-GPS technology uses a location server with a reference GPS receiver that can simultaneously detect the same satellites as the wireless handset (or mobile station) with a partial GPS receiver, to help the partial GPS receiver to find weak GPS signals. The wireless handset collects measurements from both the GPS constellation and the wireless mobile network. These measurements are combined by the location server to produce position estimation. The Locata Corporation finally invented a new precision indoor/outdoor positioning technology called Locata[15]. It is consists of a time-synchronized pseudolite transceiver called a LocataLite which transmits GPS-like signals that allow single-point positioning using carrier-phase measurements. The result shows that it is possible to archive distance error within sub-centimeter [15]. Although the ubiquitous positioning is not new, but the technique above is more focus to the new device development, rather than utilize mobility function on mobile device.

3 System Design

Generally our system is to determine positioning information inside and outside building, by using mobile device. In order to utilize mobility function in our system, the integration of internal positioning sensor (such as: GPS, GSM, WLAN and Bluetooth) within a mobile device is needed. Refer to Figure 1, our system generally consist of three (3) subsystems which are named as field subsystem, interface subsystem, and database subsystem. In the normal situation GPS satellite (in earth orbit) and WLAN access point (inside building) will continuously broadcasting their signal within coverage. Any mobile device that equipped WLAN and GPS sensor within their coverage will received signal. The signal will be processed by conventional indoor/outdoor positioning (as shown on Figure 2). This algorithm is aimed to switch suitable positioning method in a different environment. The method indoor positioning will be started as soon as outdoor positioning sensor cannot get signal input. In order to select which positioning method will be used, we are using GPS as outdoor positioning sensor, meanwhile WLAN as indoor positioning sensor. Then, the positioning data in obtained signal will be extracted to compare with surveying data in database server. Finally the output of system will display the mapping location on mobile device screen.

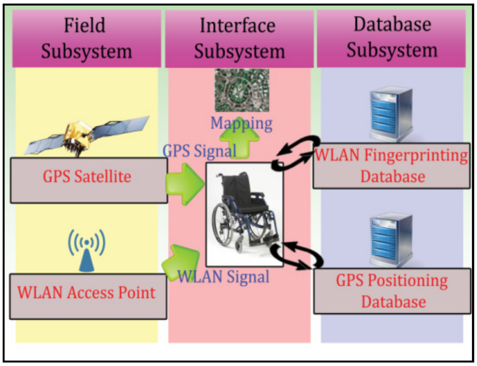


Fig. 1. General Architecture of Wheelchair Navigation

```
U-Navigation Algorithm {
1) Read GPS Signal;
2) If GPS Signal is valid {
   A. Outdoor_Positioning;
   }
3) else {
   B. Apply indoor_
      positioning;

  }
```

Fig. 2. Basic Algorithm of Wheelchair Navigation (written in Pseudocode)

3.1 Outdoor Positioning

In the GPS positioning we need to set up some known points or so called references points in order to compare within obtained coordinates from GPS signal. The reason behind this is to map the real environment within mobile device screen. Thus, we prefer to use Relative-Interpolation method [16] since the coordinates of the reference point are not the absolute longitude and latitude, but the x-y coordinates relative to the (0, 0) point of the window in which the map is rendered. At each of the reference points, A and B, we need to know the coordinate (real environment) by using high precision GPS positioning on the site (for example using GPS device in Figure 4) and, longitude and latitude as well as its X and Y coordinates on the window. Let x_{lat} and x_{long} be the longitude and latitude of the reference point x (A or B), and x_x and x_y be it is x and y coordinates, respectively. Let C be the unknown user's current position and c_{lat} and c_{long} be the GPS data measured at the user's current position. Then, we estimate the user's x and y coordinates on the map, c_x and c_y respectively, with the equations below:

$$C_x = \left(\frac{C_{lon} - A_{lon}}{B_{lon} - A_{lon}}\right)(B_x - A_x) + A_x \tag{1}$$

$$C_y = \left(\frac{C_{lat} - A_{lat}}{B_{lat} - A_{lat}}\right)(B_y - A_y) + A_y \tag{2}$$

3.2 Indoor Positioning

In our indoor positioning method, we prefer to use a well known WLAN fingerprinting method which known as RADAR [12]. In the RADAR WLAN positioning system, there is a searching algorithm, which it is be in main part of the system, known as KNN nearest neighbor [17]. This algorithm contributes by look-up table during the off-line phase. However, WLAN signal strength also suffers in the obstruction environment since the signal will propagate and loss if there is a blockage between AP and mobile device receiver. Theoretically, the WLAN signal path loss obeys the distance power law as described below;

$$P_r(d) = P_r(d_0) - 10nlog\left(\frac{d}{d_0}\right) + X_\sigma \tag{3}$$

Where Pr is the received power; $P_r(d_0)$ is the received power at the d_0(called as reference distance), n is refer to path loss exponent, which indicates the rate of the path loss increases with distance. It depends on the surrounding, building type and other obstructions. And d_0 is the close-in reference distance (1m) and d is the distance of separation between the RF signal transmitter and receiver (The transmitter could be AP and receiver could be mobile device receiver). The term X_σ is a zero mean Gaussian random variable with standard deviation σ. Equation (3) is modified to include Wall Attenuation Factor (WAF). The modified distance power law is given as (4),

$$P_r(d) = P_r(d_0) - 10nlog\left(\frac{d}{d_0}\right) - T * WAF \tag{4}$$

Where, T is number of walls between transmitter and receiver.

$$\mathrm{d} = \mathrm{e}^{\left(\frac{\mathrm{P_r(d_0)-P_r(d)-T*WAF}}{10}\right)} \tag{5}$$

Equation (5) has been derived from equation (4). This equation is to measure the distance between the Access Point and Mobile Node. When the mobile device or node location been calculated, the distance of every devices or nodes will be calculated by using Euclidean Distance equation (6).

$$\mathrm{Distance} = \sqrt{((\mathrm{X_1} - \mathrm{X_2})^2 + (\mathrm{Y_1} - \mathrm{Y_2})^2)} \tag{6}$$

The Location Server will calculate the distance for every device in the network and compare all the distance to find out which is the nearest device from the mobile node chosen. The nearest computation method is done by the nearest neighbor(s) in signal space (NNSS). The idea is to compute the distance (in signal space) between the observed set of SS measurements, $(\mathrm{ss_1}, \mathrm{ss_2}, \mathrm{ss_3})$ and the recorded SS, $(\mathrm{ss_1'}, \mathrm{ss_2'}, \mathrm{ss_3'})$ at a fixed set of locations, and then pick the location that minimizes the distance. In order to calculate based on three (3) measurements, the equation (6) can be inherit to become as equation (7) which is D is distance between observed signal and recorded signal;

$$\mathrm{D} = \sqrt{((\mathrm{ss_1} - \mathrm{ss_1'})^2 + (\mathrm{ss_2} - \mathrm{ss_2'})^2 + (\mathrm{ss_3} - \mathrm{ss_3'})^2)} \tag{7}$$

4 Performance Evaluation

We categorize our experiment in two (2) types; which are indoor positioning and outdoor positioning. These experiments were conducted at Universiti Teknologi Malaysia, Johor, Malaysia specifically; indoor positioning at FSKSM building, and outdoor positioning at Lingkaran Ilmu road (see Figure 3). At indoor positioning experiment, we required to collect WLAN signal strength at two (2) location by walking along blue path at two (2) different locations. In this part, we are using mobile device (model: HTC HD Mini, software: WiFiFofum) to collect the data in four (4) orientation for each point. Meanwhile, at outdoor positioning, the data collection was obtained by taking GPS coordinate point (using GPS Trimble device) along red path. These collected data (GPS coordinate and WLAN signal strength) will be stored at database subsystem. Finally, the positioning accuracy result can be obtained by comparing the current positioning information with positioning information which stored at database subsystem Below, we will discuss more detail about performance of proposed approach.

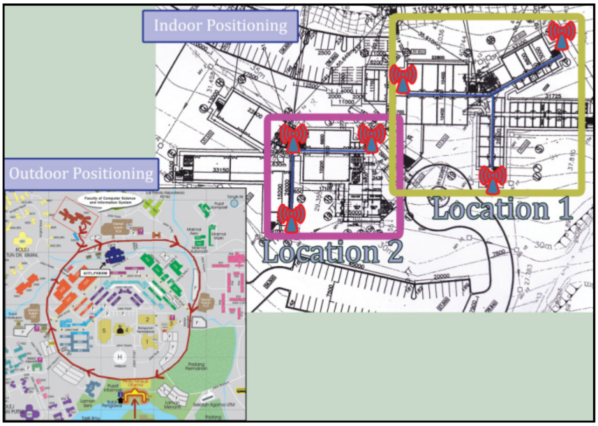


Fig. 3. Experiment Area (see red path for outdoor positioning experiment area). For Indoor Positioning, There Two Types of Location Which Known; as Location 1 and Location 2 (see blue path for indoor positioning experiment area).

4.1 Outdoor Positioning

The measured x-y coordinates and the latitude and longitude of the reference points are shown in Table 1. We performed experiments in which the x-y coordinates of the current position obtained by clicking the mouse on the window were compared with those obtained from the outdoor positioning program 200 times and the results are summarized in Table 2. The results indicate that among the 200 experiments, on 11 occasions the error was less than 1 m, on 17 occasions the error was between 1 and 2 m, and so on. The average error was 4.875m.

Table 1. x,y coordinates, latitude and longitude of reference points

	Coordinates		**GPS data**	
	X	**Y**	**Latitude**	**Longitude**
A	1842	1140	1°33'27.16" N	103°38'11.69"E
B	2112	1156	1°33'45.84"N	103°38'13.82"E

Table 2. Summary of the results of the outdoor positioning experiments

Error (m)	0~1	**1~2**	**2~4 `**	**4~6**	**6~8**	**8~**
Occurrence	11	17	61	51	33	27
Probability	5.5 %	8.5%	30.5%	25.5 %	16.5 %	13.5%
Average Error = 4.875 m						

4.2 Indoor Positioning

In the end, we plot the overall result of WLAN positioning in Figure 4 (Location 1) and Figure 5 (Location 2). According to that figure, it shows, most of the positioning information result getting error during WLAN mobile devices receiver located in the middle of Location 1. The reason behind this is due to the signal strength that obtained in that area is almost the same, and it makes the searching algorithm (KNN nearest neighbor) easier to choose unsuitable signal strength that refer to location in the database server (the signal strength that refer to the location was saved in the offline phase). However, the positioning information results are getting better during WLAN mobile devices receiver that is not located in the middle of location anymore. The result is 37.35 percent of distance error at Location 1 was below 3.4 m. Besides that, 56.67 percent of distance error at Location 2 was below 2.9m. at the Location 1. As overall result, the average error of the indoor positioning in was 7.69777193 meter (Location 1) and 6.12233997 meter (Location 2).

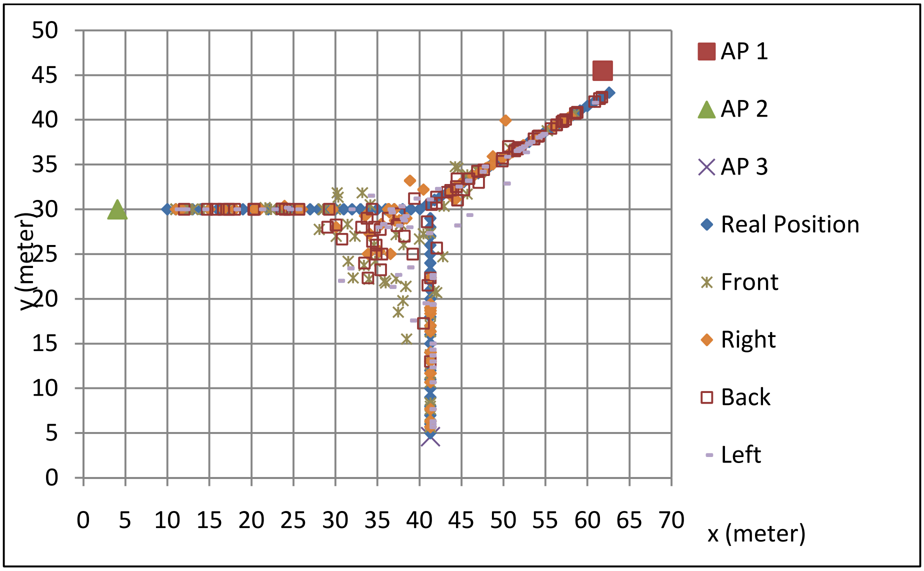


Fig. 4. Localization Comparison Between Real Position And Experimented Position at Location 1. Front, Right, Back, and Left Refer to User Orientation during Experiment.

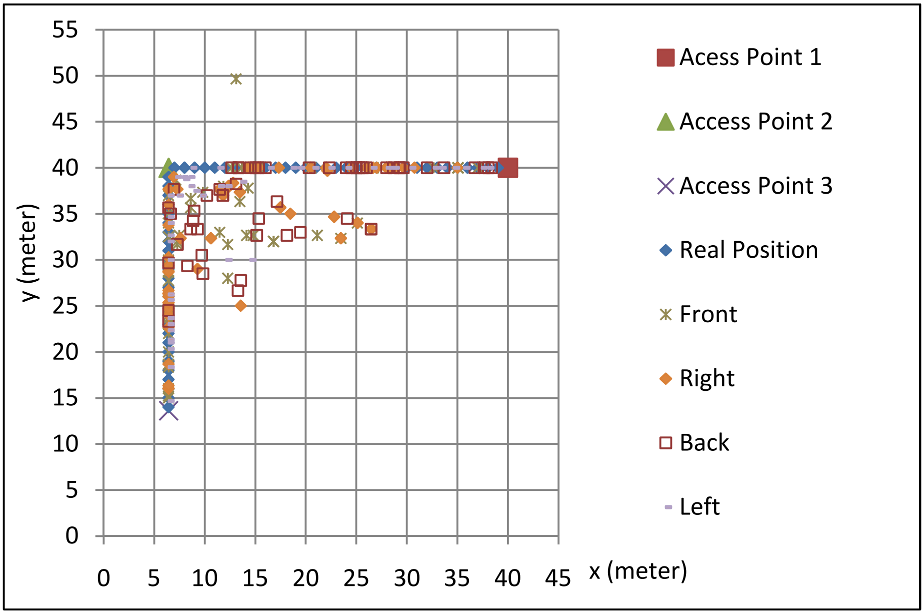


Fig. 5. Localization Comparison Between Real Position And Experimented Position at Location 2. Front, Right, Back, and Left Refer to User Orientation during Experiment.

5 Conclusion and Future Works

This paper discussed about our experiment for location determination by using integration between GPS and WLAN for wheelchair navigation system. The information from both of sensor output that was switched based on type of environment (whether inside or outside building) in order to find the absolute user target position. The result shows our proposed method can archive 7.69777193 meter (Location 1) and 6.12233997 meter (Location 2) during indoor positioning. Besides that, our wheelchair navigation system can survive during outdoor environment by archive 4.875m distance error. We believe our result is suitable and accepted for wheelchair navigation purpose, although sometime the distance error of indoor environment slightly going too large. As future works, we will continue our experiment by using this result and combine with other mobile internal sensor such as camera in order to know how far our approach can be more ubiquity in many environments.

Acknowledgement. This work is financed by Zamalah Scholarship provided by Universiti Teknologi Malaysia and the Ministry of Higher Education Malaysia.

References

[1] Benini, A., Mancini, A., Longhi, S.: An IMU/UWB/Vision-based Extended Kalman Filter for Mini-UAV Localization in Indoor Environment using 802.15. 4a Wireless Sensor Network. Journal of Intelligent & Robotic Systems, 1–16 (2012)

[2] Dhital, A., Closas, P., Fernández-Prades, C.: Bayesian filtering for indoor localization and tracking in wireless sensor networks. EURASIP Journal on Wireless Communications and Networking 2012(1), 21 (2012)

[3] Ren, M., Karimi, H.A.: Movement Pattern Recognition Assisted Map Matching for Pedestrian/Wheelchair Navigation. The Journal of Navigation 65(04), 617–633 (2012)

[4] Niitsuma, M., Ochi, T., Yamaguchi, M., Iwamot, K.: Design of Mutual Interaction Between a User and Smart Electric Wheelchair. Journal of Advanced Computational Intelligence and Intelligent Informatics 16(2), 305–312 (2012)

[5] Zhang, L., Sun, S., Huang, Y., Zhu, Y., Shi, J.: Research on obstacle avoidance system of intelligent wheelchair. Journal of Electronic Measurement and Instrument 12, 003 (2011)

[6] Boulos, M.K., Berry, G.: Real-time locating systems (RTLS) in healthcare: a condensed primer. International Journal of Health Geographics 11(1), 25 (2012)

[7] Abascal, J., Lafuente, A., Marco, A., Falco, J.M., Casas, R., Sevillano, J.L., Cascado, D., Lujan, C.: An architecture for assisted navigation in intelligent environments. International Journal of Communication Networks and Distributed Systems 4(1), 49–69 (2010)

[8] Bejuri, W.M.Y.W., Mohamad, M.M., Sapri, M.: Ubiquitous positioning: A Taxonomy for Location Determination on Mobile Navigation System. Signal & Image Processing: An International Journal (SIPIJ) 2(1), 24–34 (2011)

[9] Bejuri, W.M.Y.W., Mohamad, M.M., Sapri, M., Rosly, M.A.: Investigation of Color Constancy for Ubiquitous Wireless LAN/Camera Positioning: An Initial Outcome. International Journal of Advancements in Computing Technology 4(7), 269–280 (2012)

[10] Bejuri, W.M.Y.W., Mohamad, M.M., Sapri, M., Rosly, M.A.: Ubiquitous WLAN/Camera Positioning using Inverse Intensity Chromaticity Space-based Feature Detection and Matching: A Preliminary Result. Presented at the International Conference on Man-Machine Systems 2012, ICOMMS 2012, Penang, Malaysia (2012)

[11] Bejuri, W.M.Y.W., Mohamad, M.M., Sapri, M., Rosly, M.A.: Performance Evaluation of Mobile U-Navigation based on GPS/WLAN Hybridization. Journal of Convergence Information Technology 7(12), 235–246 (2012)

[12] Bahl, P., Padmanabhan, V.N.: RADAR: An in-building RF-based user location and tracking system. In: Proceedings of the Nineteenth Annual Joint Conference of the IEEE Computer and Communications Societies, INFOCOM 2000, vol. 2, pp. 775–784. IEEE (2000)

[13] Youssef, M.A., Agrawala, A., Udaya Shankar, A.: WLAN location determination via clustering and probability distributions. In: Proceedings of the First IEEE International Conference on Pervasive Computing and Communications, PerCom 2003, pp. 143–150 (2003)

[14] Castro, P., Chiu, P., Kremenek, T., Muntz, R.: A Probabilistic Room Location Service for Wireless Networked Environments. In: Abowd, G.D., Brumitt, B., Shafer, S.A.N. (eds.) UbiComp 2001. LNCS, vol. 2201, pp. 18–34. Springer, Heidelberg (2001)

[15] Barnes, J., Rizos, C., Wang, J., Small, D., Voigt, G., Gambale, N.: Locata: the positioning technology of the future. In: Proceedings of the 6th International Symposium on Satellite Navigation Technology Including Mobile Positioning and Location Services (July 2003)

[16] Yim, J., Ko, I., Do, J., Joo, J., Jeong, S.: Implementation of a Prototype Positioning System for LBS on U-campus. Journal of Universal Computer Science 14(14), 2381–2399 (2008)

[17] Wettschereck, D., Dietterich, T.G.: An experimental comparison of the nearest-neighbor and nearest-hyperrectangle algorithms. Machine Learning 19(1), 5–27 (1995)

Correctness of Solving Query-Answering Problems Using Satisfiability Solvers

Kiyoshi Akama[1] and Ekawit Nantajeewarawat[2]

[1] Information Initiative Center, Hokkaido University, Hokkaido, Japan
akama@iic.hokudai.ac.jp

[2] Computer Science Program, Sirindhorn International Institute of Technology
Thammasat University, Pathumthani, Thailand
ekawit@siit.tu.ac.th

Abstract. A query-answering (QA) problem is concerned with finding the set of all ground instances of a given atomic formula that are logical consequences of a specified logical formula. Recently, many kinds of problems have been solved efficiently by using satisfiability (SAT) solvers, motivating us to use SAT solvers to speed up solving a class of QA problems. Given a finite ground clause set as input, a SAT solver used in this paper generates all models of the input set that contain only atomic formulas appearing in it. A method for solving QA problems using SAT solvers is developed, based on the use of a support set to restrict the generation of ground instances of given clauses possibly with constraint atomic formulas. The correctness of the proposed method is proved.

Keywords: Query-answering problems; SAT solvers; problem solving.

1 Introduction

A *query-answering problem* (*QA problem*) is a pair $\langle K, a \rangle$, where K is a logical formula, representing background knowledge, and a is an atomic formula (atom), representing a query. The answer to a QA problem $\langle K, a \rangle$ is the set of all ground instances of a that are logical consequences of K. Several specific classes of QA problems have been discussed in previous works, e.g., QA problems on definite clauses [6] and those on description logics [10].

Given a finite input set of propositional clauses, a usual satisfiability solver (SAT solver) returns a model of the input set if the set is satisfiable. Efficient SAT solvers have been invented [4,7,9], and many kinds of problems have been solved effectively by using SAT solvers. This motivates us to use SAT solvers to speed up solving a class of QA problems. Solving QA problems, however, requires generation of all elements of an answer set, and thus, a usual SAT solver is not directly applicable. We use in this paper a class of SAT solvers that generate from a given finite input ground clause set all of its models that contain only atoms appearing in it.

The objective of this paper is to develop a method for solving QA problems on arbitrary clauses using SAT solvers and to prove the correctness of the method.

A. Selamat et al. (Eds.): ACIIDS 2013, Part I, LNAI 7802, pp. 404–413, 2013.

Let a QA problem $\langle K, a \rangle$ be given, where K is a set of clauses. A primary question needed to be addressed is how to instantiate the clause set K into a finite set of ground clauses to be input to a SAT solver. The solution provided by our method is to generate a support set from K and then instantiate K into ground clauses that contain only atoms in the support set. When the generated support set is finite, the set of obtained ground clauses is also finite, yielding a SAT-solver input clause set.

To begin with, Section 2 formalizes QA problems. After introducing the notion of a representative set, Section 3 defines a representative set of the collection of all models of a set of clauses, which provides a basis for constructing a support set. Section 4 presents our proposed method and Section 5 proves its correctness. Section 6 illustrates its application. Section 7 provides conclusions.

2 Query-Answering Problems and Model Intersection

2.1 Query-Answering (QA) Problems

A *query-answering problem* (*QA problem*) is a pair $\langle K, a \rangle$, where K is a logical formula and a is an atomic formula (atom). The *answer* to a QA problem $\langle K, a \rangle$, denoted by $ans(K, a)$, is defined by

$$ans(K, a) \;=\; \{a' \mid (a' \text{ is a ground instance of } a) \;\&\; (K \models a')\},$$

i.e., the set of all ground instances of a that follow logically from K. When K consists of only definite clauses, problems in this class are problems that have been discussed in logic programming [6]. In the class of QA problems discussed in [10], K is a conjunction of axioms and assertions in Description Logics [3]. Recently, QA problems have gained wide attention, owing partly to emerging applications in systems involving integration between ontological knowledge and instance-level rule-oriented components, e.g., interaction between Description Logics and Horn rules [5,8] in the Semantic Web's ontology-based rule layer.

In this paper, we focus our attention on the class of QA problems on arbitrary clauses possibly containing constraint atoms. A QA problem in this class takes the form $\langle K, a \rangle$, where (i) K is a set of clauses, each of which is an expression $(a_1, \ldots, a_m \leftarrow b_1, \ldots, b_n)$ such that $m \geq 0$, $n \geq 0$, and each of $a_1, \ldots, a_m$, $b_1, \ldots, b_n$ is either a usual atom or a constraint atom, and (ii) a is a usual atom. Unlike usual atoms, the truth values of ground instances of constraint atoms are predetermined independently of interpretations.

2.2 QA Problems as Model Intersection Problems

In the rest of this paper, let $\mathcal{G}$ denote the set of all ground usual atoms. An *interpretation* is a subset of $\mathcal{G}$. Assume that K is a first-order formula. A *model* of K is an interpretation that satisfies K. Using the set of all models of K, denoted by $Models(K)$, the answer to a QA problem $\langle K, a \rangle$ can be equivalently represented as

$$ans(K, a) \;=\; (\bigcap Models(K)) \cap rep(a),$$

where $\bigcap \mathit{Models}(K)$ is the intersection of all models of K and $\mathit{rep}(a)$ is the set of all ground instances of a.

3 A Representative Set for Solving QA Problems

Next, the notion of a representative set of a collection of sets is introduced. The intersection of a given collection of sets can be determined in terms of the intersection of sets in its representative set (Theorem 1). Given a set Cs of clauses, a set collection, $\mathbb{MM}(\mathit{Cs})$, is defined, with an important property being that $\mathbb{MM}(\mathit{Cs})$ is a representative set of the set of all models of Cs (Theorem 2). Consequently, the answer to a QA problem concerning Cs can be computed through $\mathbb{MM}(\mathit{Cs})$.

3.1 Representative Sets

For any set G, let $\mathit{pow}(G)$ denote the power set of G. A representative set is defined below:

Definition 1. Let G be a set and $M_1, M_2 \subseteq \mathit{pow}(G)$. M_1 is a *representative set* of M_2 iff the following conditions are satisfied:

1. $M_1 \subseteq M_2$.
2. For any $m_2 \in M_2$, there exists $m_1 \in M_1$ such that $m_2 \supseteq m_1$. □

Theorem 1 below provides a basis for computing the intersection of the set of all models of a clause set using its representative set. Its proof is given in [2].

Theorem 1. *Let G be a set and $M_1, M_2 \subseteq \mathit{pow}(G)$. If M_1 is a representative set of M_2, then $\bigcap M_1 = \bigcap M_2$.* □

3.2 A Representative Set for All Models of a Definite-Clause Set

A *definite clause* is a clause whose left-hand side contains exactly one usual atom and no constraint atom. Given a definite clause C, the atom in the left-hand side of C is called the *head* of C, denoted by $\mathit{head}(C)$, and the set of all usual atoms and constraint atoms in the right-hand side of C is called the *body* of C, denoted by $\mathit{body}(C)$. Assume that D is a set of definite clauses possibly with constraint atoms. The *meaning* of D, denoted by $\mathcal{M}(D)$, is defined as follows:

1. A mapping T_D on $\mathit{pow}(\mathcal{G})$ is defined by: for any set $G \subseteq \mathcal{G}$, $T_D(G)$ is the set
$$\begin{aligned}\{\mathit{head}(C\theta) \mid\ & (C \in D)\ \&\\ & (\theta \text{ is a ground substitution for all variables occurring in } C)\ \&\\ & (\text{each usual atom in } \mathit{body}(C\theta) \text{ belongs to } G)\ \&\\ & (\text{each constraint atom in } \mathit{body}(C\theta) \text{ is true})\}.\end{aligned}$$
2. $\mathcal{M}(D)$ is then defined as the set $\bigcup_{n=1}^{\infty} T_D^n(\varnothing)$, where $T_D^1(\varnothing) = T_D(\varnothing)$ and $T_D^n(\varnothing) = T_D(T_D^{n-1}(\varnothing))$ for each $n > 1$.

It is well known that $\mathcal{M}(D)$ is the least model of D. So the singleton set $\{\mathcal{M}(D)\}$ is a representative set of $\mathit{Models}(D)$.

3.3 A Representative Set for All Models of a Set of Clauses

Given a clause C, the set of all usual atoms and constraint atoms in the left-hand side of C is denoted by $\mathit{lhs}(C)$ and the set of all those in the right-hand side of C is denoted by $\mathit{rhs}(C)$. It is assumed henceforth that: (i) for any constraint atom a, $\mathit{not}(a)$ is a constraint atom; (ii) for any constraint atom a and any substitution θ, $\mathit{not}(a)\theta = \mathit{not}(a\theta)$; and (iii) for any ground constraint atom a, a is true iff $\mathit{not}(a)$ is not true.

The following notation is used for defining a representative set of the set of all models of a given clause set (Definition 2).

1. Let Cs be a set of clauses possibly with constraint atoms.
   (a) $\textsc{mvRhs}(\mathit{Cs})$ is defined as the set $\{\textsc{mvRhs}(C) \mid C \in \mathit{Cs}\}$, where for any clause $C \in \mathit{Cs}$, $\textsc{mvRhs}(C)$ is the clause obtained from C as follows: For each constraint atom c in $\mathit{lhs}(C)$, remove c from $\mathit{lhs}(C)$ and add $\mathit{not}(c)$ to $\mathit{rhs}(C)$.
   (b) For any $G \subseteq \mathcal{G}$, $\textsc{Ginst}(\mathit{Cs}, G)$ is defined as the set

   $$\begin{aligned}\{C\theta \mid\ & (C \in \mathit{Cs})\ \&\\ & (\theta \text{ is a ground substitution for all variables occurring in } C)\ \&\\ & (\text{each usual atom in } C\theta \text{ belongs to } G)\}.\end{aligned}$$

   (c) For any $G \subseteq \mathcal{G}$, $\textsc{Inst}(\mathit{Cs}, G)$ is defined by

   $$\textsc{Inst}(\mathit{Cs}, G) \ = \ \textsc{Ginst}(\textsc{mvRhs}(\mathit{Cs}), G).$$

2. Let Cs be a set of ground clauses with no constraint atom in their left-hand sides. We can construct a set of definite clauses from Cs as follows: For each clause $C \in \mathit{Cs}$,
   - if $\mathit{lhs}(C) = \varnothing$, then construct a definite clause the head of which is $\perp$ and the body of which is $\mathit{rhs}(C)$, where $\perp$ is a special symbol not occurring in Cs;
   - if $\mathit{lhs}(C) \neq \varnothing$, then (i) select one arbitrary atom a from $\mathit{lhs}(C)$, and (ii) construct a definite clause the head of which is a and the body of which is $\mathit{rhs}(C)$.

   Let $\textsc{Dc}(\mathit{Cs})$ denote the set of all definite-clause sets possibly constructed from Cs in the above way.

Now let Cs be a set of clauses possibly with constraint atoms. A representative set, $\mathbb{MM}(\mathit{Cs})$, of $\mathit{Models}(\mathit{Cs})$ is defined below.

Definition 2. A set $\mathbb{MM}(\mathit{Cs})$ of ground-atom sets is defined by

$$\mathbb{MM}(\mathit{Cs}) \ = \ \{\mathcal{M}(D) \mid (D \in \textsc{Dc}(\textsc{Inst}(\mathit{Cs}, \mathcal{G})))\ \&\ (\perp \notin \mathcal{M}(D))\}. \qquad \square$$

It is shown in [2] that:

Theorem 2. *$\mathbb{MM}(\mathit{Cs})$ is a representative set of $\mathit{Models}(\mathit{Cs})$.* □

Along with Theorem 1, Theorem 2 shows that $\bigcap \mathbb{MM}(\mathit{Cs}) = \bigcap \mathit{Models}(\mathit{Cs})$. As a result, for any usual atom a, $\mathit{ans}(\mathit{Cs}, a) = (\bigcap \mathbb{MM}(\mathit{Cs})) \cap \mathit{rep}(a)$.

4 Solving QA Problems with SAT Solvers

After introducing the notion of a support set (Definition 3) and its construction (Theorem 3), a method for solving QA problems using SAT solvers is described (Section 4.2).

4.1 Support Sets and Construction of Support Sets

Let Cs be a set of clauses possibly with constraint atoms. A support set for Cs is defined as follows:

Definition 3. A *support set* for Cs is a set $\mathbb{S}$ of ground usual atoms such that $\bigcup \mathbb{MM}(Cs) \subseteq \mathbb{S}$. □

Proposition 1 and Theorem 3 below illuminate how to construct a support set for Cs. Given a clause $C \in Cs$, let $split(C)$ and $split_{\perp}(C)$ be defined as follows:

1. $split(C)$ is the set of all definite clauses C' such that $head(C') \in lhs(C)$ and $body(C') = rhs(C)$.
2. If $lhs(C) = \varnothing$, then $split_{\perp}(C) = \{C'\}$, where C' is the definite clause such that $head(C') = \perp$ and $body(C') = rhs(C)$.
3. If $lhs(C) \neq \varnothing$, then $split_{\perp}(C) = split(C)$.

Proposition 1. *For any* $D \in \text{Dc}(\text{Inst}(Cs, \mathcal{G}))$,

$$\mathcal{M}(D) \subseteq \mathcal{M}(\bigcup\{split_{\perp}(C) \mid C \in \text{mvRhs}(Cs)\}).$$

Proof. Let $D \in \text{Dc}(\text{Inst}(Cs, \mathcal{G}))$. Assume that

- $\hat{D}_1 = \bigcup\{split_{\perp}(C) \mid C \in \text{Inst}(Cs, \mathcal{G})\}$,
- $\hat{D}_2 = \text{Ginst}(\bigcup\{split_{\perp}(C) \mid C \in \text{mvRhs}(Cs)\}, \mathcal{G})$, and
- $\hat{D}_3 = \bigcup\{split_{\perp}(C) \mid C \in \text{mvRhs}(Cs)\}$.

Since $D \subseteq \hat{D}_1$, $\mathcal{M}(D) \subseteq \mathcal{M}(\hat{D}_1)$. Since $\hat{D}_1 = \hat{D}_2$, $\mathcal{M}(\hat{D}_1) = \mathcal{M}(\hat{D}_2)$. By the definition of $\mathcal{M}$, $\mathcal{M}(\hat{D}_2) = \mathcal{M}(\hat{D}_3)$. Thus $\mathcal{M}(D) \subseteq \mathcal{M}(\hat{D}_3)$. □

Theorem 3. $\mathcal{M}(\bigcup\{split(C) \mid C \in \text{mvRhs}(Cs)\})$ *is a support set for* Cs.

Proof. Assume that $\hat{D}_{\perp} = \bigcup\{split_{\perp}(C) \mid C \in \text{mvRhs}(Cs)\}$ and $\hat{D} = \bigcup\{split(C) \mid C \in \text{mvRhs}(Cs)\}$. Let $G \in \mathbb{MM}(Cs)$. Then there exists $D \in \text{Dc}(\text{Inst}(Cs, \mathcal{G}))$ such that $\perp \notin \mathcal{M}(D)$ and $G = \mathcal{M}(D)$. By Proposition 1, $\mathcal{M}(D) \subseteq \mathcal{M}(\hat{D}_{\perp})$, whence $G \subseteq \mathcal{M}(\hat{D}_{\perp})$. It follows that $\bigcup \mathbb{MM}(Cs) \subseteq \mathcal{M}(\hat{D}_{\perp})$. Since $\bigcup \mathbb{MM}(Cs)$ does not contain $\perp$, $\bigcup \mathbb{MM}(Cs) \subseteq \mathcal{M}(\hat{D}_{\perp}) - \{\perp\} = \mathcal{M}(\hat{D})$. □

4.2 A Method for Solving QA Problems Using SAT Solvers

Next, a SAT solver used in this paper is introduced. Let $\hat{Cs}$ be a given finite set of ground clauses with no occurrence of any constraint atom. For any clause

$C \in \hat{Cs}$, let *atoms*(C) denote the set *lhs*$(C) \cup$ *rhs*(C). Let *atoms*$(\hat{Cs})$ denote the set $\bigcup\{atoms(C) \mid C \in \hat{Cs}\}$. Let $\textsc{SatSol}(\hat{Cs})$ denote the set of all subsets of *atoms*$(\hat{Cs})$ that are models of $\hat{Cs}$. Given $\hat{Cs}$ as input, a *satisfiability solver* (*SAT solver*) outputs $\textsc{SatSol}(\hat{Cs})$.

The following notation is used in our proposed method for solving QA problems: Given a set $\tilde{Cs}$ of ground clauses containing no constraint atom in their left-hand sides, let $\textsc{Eval}(\tilde{Cs})$ denote the set

$$\{\textsc{rmCon}(C) \mid (C \in \tilde{Cs}) \ \& \ (\text{each constraint atom in } rhs(C) \text{ is true})\},$$

where for any clause $C \in \tilde{Cs}$, $\textsc{rmCon}(C)$ is the clause obtained from C by removing all constraint atoms from its right-hand side.

Using SAT solvers, the proposed method for solving QA problems is presented below. Let a QA problem $\langle Cs, a \rangle$ be given as input, where Cs is a set of clauses possibly containing constraint atoms.

1. Construct $\hat{D} = \bigcup\{split(C) \mid C \in \textsc{mvRhs}(Cs)\}$.
2. Let $\mathbb{S} = \mathcal{M}(\hat{D})$. By Theorem 3, $\mathbb{S}$ is a support set for Cs.
3. If $\mathbb{S}$ is not finite, then stop with failure.
4. Construct $\hat{Cs} = \textsc{Eval}(\textsc{Inst}(Cs, \mathbb{S}))$. Since $\mathbb{S}$ is finite, $\hat{Cs}$ is a finite set of ground clauses.
5. Solve $\hat{Cs}$ by a SAT solver to obtain $\textsc{SatSol}(\hat{Cs})$.
6. Output $(\bigcap \textsc{SatSol}(\hat{Cs})) \cap rep(a)$ as the answer to the QA problem $\langle Cs, a \rangle$.

5 Correctness of the Proposed Method

Using basic results established in Section 5.1, the correctness of the proposed method is shown in Section 5.2 (Theorem 5).

5.1 Ground Clauses Used for Bottom-Up Computation

Let D be a set of ground definite clauses with no occurrence of any constraint atom. Associated with D is a mapping $\hat{T}_D$ on $pow(\mathcal{G}) \times pow(D)$ defined as follows: For any $G \subseteq \mathcal{G}$ and any $R \subseteq D$, $\hat{T}_D(G, R) = \langle G', R' \rangle$, where

1. $G' = \{head(C) \mid (C \in D) \ \& \ (body(C) \subseteq G)\}$,
2. $R' = \{C \mid (C \in D) \ \& \ (body(C) \subseteq G)\}$.

It is obvious that:

Proposition 2. *Assume that* $G_1, G_1', G_2, G_2' \subseteq \mathcal{G}$ *and* $R_1, R_1', R_2, R_2' \subseteq D$ *such that* $\hat{T}_D(G_1, R_1) = \langle G_1', R_1' \rangle$ *and* $\hat{T}_D(G_2, R_2) = \langle G_2', R_2' \rangle$. *If* $G_1 \subseteq G_2$, *then* $G_1' \subseteq G_2'$ *and* $R_1' \subseteq R_2'$. □

Next, let a sequence $Seq(D) = \langle G_0, R_0 \rangle, \langle G_1, R_1 \rangle, \langle G_2, R_2 \rangle, \ldots$ be defined by: $G_0 = \varnothing$, $R_0 = \varnothing$, and for any integer $i \geq 0$, $\langle G_{i+1}, R_{i+1} \rangle = \hat{T}_D(G_i, R_i)$. Let G_∞ and R_∞ be defined by: $G_\infty = \bigcup_{i=1}^{\infty} G_i$ and $R_\infty = \bigcup_{i=1}^{\infty} R_i$.

Proposition 3. *For any integer $i \geq 0$, $G_i \subseteq G_{i+1}$ and $R_i \subseteq R_{i+1}$.*

Proof. The result is proved by induction on i. Obviously, $\varnothing = G_0 \subseteq G_1$ and $\varnothing = R_0 \subseteq R_1$. Let $i \geq 0$ and assume that $G_i \subseteq G_{i+1}$. By Proposition 2, $G_{i+1} \subseteq G_{i+2}$ and $R_{i+1} \subseteq R_{i+2}$. □

Proposition 4. *For any integer $i \geq 0$, $G_i = \mathit{atoms}(R_i)$.*

Proof. Obviously $\varnothing = G_0 = \mathit{atoms}(R_0)$. Now let $i > 0$. First we show that $G_i \subseteq \mathit{atoms}(R_i)$. Let $g \in G_i$. Then there exists $C \in D$ such that $\mathit{head}(C) = g$ and $\mathit{body}(C) \subseteq G_{i-1}$. So $C \in R_i$, whence $g \in \mathit{atoms}(R_i)$. Next, we show that $G_i \supseteq \mathit{atoms}(R_i)$. Let $C \in R_i$. Then $\mathit{body}(C) \subseteq G_{i-1}$, and thus $\mathit{head}(C) \in G_i$. By Proposition 3, $G_{i-1} \subseteq G_i$, whence $\mathit{body}(C) \subseteq G_i$. Then $\mathit{atoms}(C) \subseteq G_i$. It follows that $\mathit{atoms}(R_i) \subseteq G_i$. □

Proposition 5. *$G_\infty = \mathcal{M}(D) = \mathit{atoms}(R_\infty)$.*

Proof. By the definitions of $\mathcal{M}(D)$ and G_∞, it is obvious that $\mathcal{M}(D) = G_\infty$. By Proposition 4, $G_\infty = \mathit{atoms}(R_\infty)$. □

5.2 Correctness of Solutions with SAT Solvers

Theorem 4 below is used along with Theorems 1 and 2 for proving the main correctness theorem (Theorem 5).

Theorem 4. *Let Cs be a set of clauses possibly with constraint atoms. For any support set $\mathbb{S}$ for Cs, if $\hat{Cs} = \textsc{Eval}(\textsc{Inst}(Cs, \mathbb{S}))$, then $\mathbb{MM}(Cs)$ is a representative set of $\textsc{SatSol}(\hat{Cs})$.*

Proof. Let $\mathbb{S}$ be a support set for Cs and $\hat{Cs} = \textsc{Eval}(\textsc{Inst}(Cs, \mathbb{S}))$. First, we show that $\mathbb{MM}(Cs) \subseteq \textsc{SatSol}(\hat{Cs})$. Assume that $G \in \mathbb{MM}(Cs)$. Then $G = \mathcal{M}(D)$ for some $D \in \textsc{Dc}(\textsc{Inst}(Cs, \mathcal{G}))$ such that $\bot \notin \mathcal{M}(D)$. Since $\mathbb{S}$ is a support set for Cs, $G \subseteq \mathbb{S}$. To prove that $G \in \textsc{SatSol}(\hat{Cs})$, we show that $G \subseteq \mathit{atoms}(\hat{Cs})$ and G is a model of $\hat{Cs}$ as follows:

- Let $D' = \textsc{Eval}(D)$. Obviously, $\mathcal{M}(D') = \mathcal{M}(D)$. Let R_∞ be obtained from D' using $\mathit{Seq}(D')$. By Proposition 5, $\mathcal{M}(D') = \mathit{atoms}(R_\infty)$. Since $\bot \notin \mathcal{M}(D)$, R_∞ contains no definite clause whose head is $\bot$. Then, since $G \subseteq \mathbb{S}$, each definite clause in R_∞ is obtained from some clause in $\hat{Cs}$ by deleting zero or more atoms in its left-hand side. So $\mathit{atoms}(R_\infty) \subseteq \mathit{atoms}(\hat{Cs})$. It follows that $G = \mathcal{M}(D) = \mathcal{M}(D') = \mathit{atoms}(R_\infty) \subseteq \mathit{atoms}(\hat{Cs})$.
- D is obtained from Cs by selecting one atom from the left-hand side of each ground instance of each positive clause in Cs and adding $\bot$ to the left-hand side of each ground instance of each negative clause in Cs. Then G satisfies all positive clauses in Cs. Since $\bot \notin \mathcal{M}(D)$, G satisfies all negative clauses in Cs. Since $\hat{Cs} = \textsc{Eval}(\textsc{Inst}(Cs, \mathbb{S})) \subseteq \textsc{Eval}(\textsc{Inst}(Cs, \mathcal{G}))$, G satisfies all clauses in $\hat{Cs}$. So G is a model of $\hat{Cs}$.

Next, we prove that for any $G \in \textsc{SatSol}(\hat{Cs})$, there exists an atom set G' such that $G' \in \mathbb{MM}(Cs)$ and $G' \subseteq G$. Assume that $G \in \textsc{SatSol}(\hat{Cs})$. Let $G' = \mathcal{M}(D)$, where $D = \{\delta(C) \mid C \in \textsc{Inst}(Cs, \mathcal{G})\}$ and for each clause $C \in \textsc{Inst}(Cs, \mathcal{G})$, $\delta(C)$ is a definite clause defined as follows:

1. If all constraint atoms in $rhs(C)$ are true and $rhs(\textsc{rmCon}(C)) \subseteq G$, then select $head(\delta(C))$ from $lhs(C) \cap G$.
2. If some constraint atom in $rhs(C)$ is false or $rhs(\textsc{rmCon}(C)) \nsubseteq G$, then:
   (a) If $lhs(C) = \varnothing$, then let $head(\delta(C)) = \bot$.
   (b) If $lhs(C) \neq \varnothing$, then select $head(\delta(C))$ from $lhs(C)$.
3. Let $body(\delta(C)) = rhs(C)$.

Given $C \in \textsc{Inst}(Cs, \mathcal{G})$, we next show that $\delta(C)$ is well defined. Assuming that all constraint atoms in $rhs(C)$ are true and $rhs(\textsc{rmCon}(C)) \subseteq G$, we show that $lhs(C) \cap G \neq \varnothing$ as follows:

- First, we show that $lhs(C) \subseteq \mathbb{S}$. Assume the contrary, i.e., there exists $a \in lhs(C)$ such that $a \notin \mathbb{S}$. Then $a \in \bigcup \mathbb{MM}(Cs)$. Since $\mathbb{S}$ is a support set for Cs, a belongs to $\mathbb{S}$, which is a contradiction.
- Since $rhs(\textsc{rmCon}(C)) \subseteq G$ and $G \subseteq atoms(\hat{Cs}) \subseteq \mathbb{S}$, $rhs(\textsc{rmCon}(C))$ is included by $\mathbb{S}$. Since $lhs(C) \subseteq \mathbb{S}$, C belongs to $\textsc{Inst}(Cs, \mathbb{S})$. Since G is a model of $\hat{Cs} = \textsc{Eval}(\textsc{Inst}(Cs, \mathbb{S}))$ and G satisfies the right-hand side of C, $lhs(C) \cap G \neq \varnothing$.

By the construction of D, it is obvious that $D \in \textsc{Dc}(\textsc{Inst}(Cs, \mathcal{G}))$. Since each clause in $\textsc{Inst}(Cs, \mathcal{G})$ is true with respect to G, G is a model of D. Since $\mathcal{M}(D)$ is the least model of D, $\mathcal{M}(D) \subseteq G$. Since $\bot \notin G$, $\mathcal{M}(D)$ does not contain $\bot$. Since $D \in \textsc{Dc}(\textsc{Inst}(Cs, \mathcal{G}))$ and $\bot \notin \mathcal{M}(D)$, $\mathcal{M}(D)$ belongs to $\mathbb{MM}(Cs)$. □

Theorem 5. *Let Cs be a set of clauses possibly with constraint atoms. For any usual atom a and any finite support set $\mathbb{S}$ for Cs,*

$$ans(Cs, a) \;=\; (\bigcap \textsc{SatSol}(\textsc{Eval}(\textsc{Inst}(Cs, \mathbb{S})))) \cap rep(a).$$

Proof. By Theorem 2, $\mathbb{MM}(Cs)$ is a representative set of $Models(Cs)$. It follows from Theorem 1 that $\bigcap \mathbb{MM}(Cs) = \bigcap Models(Cs)$. Let $\mathbb{S}$ be a finite support set for Cs, and let $\hat{Cs} = \textsc{Eval}(\textsc{Inst}(Cs, \mathbb{S}))$. By Theorem 4, $\mathbb{MM}(Cs)$ is also a representative set of $\textsc{SatSol}(\hat{Cs})$. Then, by Theorem 1, $\bigcap \mathbb{MM}(Cs) = \bigcap \textsc{SatSol}(\hat{Cs})$. So $\bigcap Models(Cs) = \bigcap \textsc{SatSol}(\hat{Cs})$. Thus $ans(Cs, a) = (\bigcap \textsc{SatSol}(\hat{Cs})) \cap rep(a)$ for any usual atom a. □

6 Examples

Two examples illustrating application of the proposed method are given below.

Example 1. Assume that p, q, r, s, and t are predicate symbols for usual atoms, eq is a predicate symbol for constraint atoms, and for any ground terms t_1 and t_2, the constraint atom $eq(t_1, t_2)$ is true iff $t_1 = t_2$. Consider a QA problem $\langle Cs, q(x) \rangle$, where Cs consists of the following four clauses:

$$\begin{array}{llll} C_1\colon & q(x), s(x) \leftarrow r(x) & C_2\colon & t(x), p(y) \leftarrow \mathit{eq}(x,2), \mathit{eq}(y,2) \\ C_3\colon & q(x) \leftarrow t(x) & C_4\colon & q(x) \leftarrow p(x) \end{array}$$

To solve this QA problem, the proposed method is applied as follows: First, construct $\hat{D} = \bigcup\{\mathit{split}(C) \mid C \in \textsc{mvRhs}(\mathit{Cs})\} = \{C_1', C_1'', C_2', C_2'', C_3, C_4\}$, where

$$\begin{array}{ll} C_1' = (q(x) \leftarrow r(x)), & C_1'' = (s(x) \leftarrow r(x)), \\ C_2' = (t(x) \leftarrow \mathit{eq}(x,2), \mathit{eq}(y,2)), & C_2'' = (p(y) \leftarrow \mathit{eq}(x,2), \mathit{eq}(y,2)). \end{array}$$

Next, compute $\mathcal{M}(\hat{D}) = \{t(2), p(2), q(2)\}$. Let $\mathbb{S} = \mathcal{M}(\hat{D})$. Construct $\hat{\mathit{Cs}} = \textsc{Eval}(\textsc{Inst}(\mathit{Cs}, \mathbb{S})) = \{(t(2), p(2) \leftarrow), (q(2) \leftarrow t(2)), (q(2) \leftarrow p(2))\}$. Then solve $\hat{\mathit{Cs}}$ by a SAT solver to obtain

$$\textsc{SatSol}(\hat{\mathit{Cs}}) \;=\; \{\{t(2), q(2)\}, \{p(2), q(2)\}, \{t(2), p(2), q(2)\}\}.$$

Finally, the answer set $\mathit{ans}(\mathit{Cs}, q(x)) = (\bigcap \textsc{SatSol}(\hat{\mathit{Cs}})) \cap \mathit{rep}(q(x)) = \{q(2)\}$ is obtained. □

Example 2. Consider the Oedipus problem described in [3]. Oedipus killed his father, married his mother Iokaste, and had children with her, among them Polyneikes. Polyneikes also had children, among them Thersandros, who is not a patricide. The problem is to find a person who has a patricide child who has a non-patricide child. The difficulty of this problem arises due to the absence of information as to whether Polyneikes is a patricide or not.

Assume that (i) "*oe*," "*io*," "*po*," and "*th*" stand, respectively, for Oedipus, Iokaste, Polyneikes, and Thersandros, (ii) for any terms t_1, t_2, $\mathit{par}(t_1, t_2)$ denotes "t_1 is a parent of t_2," and (iii) for any term t, $\mathit{pat}(t)$ denotes "t is a patricide child." This problem is then represented as a QA problem $\langle \mathit{Cs}, \mathit{prob}(x)\rangle$, where *Cs* consists of the following seven clauses:

$$\begin{array}{llllll} C_1\colon & \mathit{par}(\mathit{io}, \mathit{oe}) \leftarrow & C_2\colon & \mathit{par}(\mathit{io}, \mathit{po}) \leftarrow & C_3\colon & \mathit{par}(\mathit{oe}, \mathit{po}) \leftarrow \\ C_4\colon & \mathit{par}(\mathit{po}, \mathit{th}) \leftarrow & C_5\colon & \mathit{pat}(\mathit{oe}) \leftarrow & C_6\colon & \leftarrow \mathit{pat}(\mathit{th}) \\ C_7\colon & \multicolumn{5}{l}{\mathit{prob}(x), \mathit{pat}(z) \leftarrow \mathit{par}(x,y), \mathit{par}(y,z), \mathit{pat}(y)} \end{array}$$

The problem is solved using the proposed method as follows: First, construct $\hat{D} = \bigcup\{\mathit{split}(C) \mid C \in \textsc{mvRhs}(\mathit{Cs})\} = \{C_1, C_2, C_3, C_4, C_5, C_8, C_9\}$, where

$$\begin{array}{l} C_8 = (\mathit{prob}(x) \leftarrow \mathit{par}(x,y), \mathit{par}(y,z), \mathit{pat}(y)), \\ C_9 = (\mathit{pat}(z) \leftarrow \mathit{par}(x,y), \mathit{par}(y,z), \mathit{pat}(y)). \end{array}$$

Compute $\mathcal{M}(\hat{D}) = \{\mathit{par}(\mathit{io}, \mathit{oe}), \mathit{par}(\mathit{io}, \mathit{po}), \mathit{par}(\mathit{oe}, \mathit{po}), \mathit{par}(\mathit{po}, \mathit{th}), \mathit{pat}(\mathit{oe}), \mathit{pat}(\mathit{po}), \mathit{pat}(\mathit{th}), \mathit{prob}(\mathit{io}), \mathit{prob}(\mathit{oe})\}$. Let $\mathbb{S} = \mathcal{M}(\hat{D})$. Construct $\hat{\mathit{Cs}} = \textsc{Eval}(\textsc{Inst}(\mathit{Cs}, \mathbb{S})) = \{C_1, C_2, C_3, C_4, C_5, C_6, C_{10}, C_{11}, C_{12}\}$, where

$$\begin{array}{l} C_{10} = (\mathit{prob}(\mathit{io}), \mathit{pat}(\mathit{po}) \leftarrow \mathit{par}(\mathit{io}, \mathit{oe}), \mathit{par}(\mathit{oe}, \mathit{po}), \mathit{pat}(\mathit{oe})), \\ C_{11} = (\mathit{prob}(\mathit{io}), \mathit{pat}(\mathit{th}) \leftarrow \mathit{par}(\mathit{io}, \mathit{po}), \mathit{par}(\mathit{po}, \mathit{th}), \mathit{pat}(\mathit{po})), \\ C_{12} = (\mathit{prob}(\mathit{oe}), \mathit{pat}(\mathit{th}) \leftarrow \mathit{par}(\mathit{oe}, \mathit{po}), \mathit{par}(\mathit{po}, \mathit{th}), \mathit{pat}(\mathit{po})). \end{array}$$

Applying a SAT solver to $\hat{\mathit{Cs}}$ yields

$$\textsc{SatSol}(\hat{\mathit{Cs}}) \;=\; \{A, A \cup \{\mathit{prob}(\mathit{oe})\}, A \cup \{\mathit{prob}(\mathit{oe}), \mathit{pat}(\mathit{po})\}\},$$

where $A = \{par(io, oe), par(io, po), par(oe, po), par(po, th), pat(oe), prob(io)\}$. The answer set $ans(Cs, prob(x)) = (\bigcap \textsc{SatSol}(\hat{Cs})) \cap rep(prob(x)) = \{prob(io)\}$ is then derived, i.e., Iokaste is the answer to this problem (no matter whether Polyneikes is a patricide). □

7 Conclusions

A method for solving QA problems using SAT solvers is proposed. The notion of a support set is introduced in order to appropriately restrict instantiations of a given clause set. When the set of ground instances obtained by such restriction is input to a SAT solver, the models generated by the solver and the set of all models of the original clause set have a common representative set (Theorems 2 and 4), forming a basis for proving the correctness of the method (Theorem 5). Construction of a support set is one important step in the proposed method. A method for such construction is presented (Theorem 3). Further work includes an extension of the proposed method to deal with QA problems on full first-order formulas based on meaning-preserving Skolemization [1], which requires appropriate techniques for handling function variables.

References

1. Akama, K., Nantajeewarawat, E.: Meaning-Preserving Skolemization. In: 2011 International Conference on Knowledge Engineering and Ontology Development, Paris, France, pp. 322–327 (2011)
2. Akama, K., Nantajeewarawat, E.: A Bottom-Up Algorithm for Solving Query-Answering Problems. In: Abd Manaf, A., Zeki, A., Zamani, M., Chuprat, S., El-Qawasmeh, E. (eds.) ICIEIS 2011. CCIS, vol. 252, pp. 299–313. Springer, Heidelberg (2011)
3. Baader, F., Calvanese, D., McGuinness, D.L., Nardi, D., Patel-Schneider, P.F.: The Description Logic Handbook, 2nd edn. Cambridge University Press (2007)
4. Bayardo, R.J., Schrag, R.C.: Using CSP Look-Back Techniques to Solve Real-World SAT Instances. In: 14th National Conference on Artificial Intelligence, pp. 203–208, Providence, Rhode Island (1997)
5. Horrocks, I., Patel-schneider, P.F., Bechhofer, S., Tsarkov, D.: OWL Rules: A Proposal and Prototype Implementation. Journal of Web Semantics 3, 23–40 (2005)
6. Lloyd, J.W.: Foundations of Logic Programming, 2nd edn. Springer (1987)
7. Moskewicz, M.W., Madigan, C.F., Zhao, Y., Zhang, L., Malik, S.: Chaff: Engineering an Efficient SAT Solver. In: 38th Annual Design Automation Conference, Las Vegas, Nevada, pp. 530–535 (2001)
8. Motik, B., Sattler, U., Studer, R.: Query Answering for OWL-DL with Rules. Journal of Web Semantics 3, 41–60 (2005)
9. Marques-Silva, J.P., Sakallah, K.A.: GRASP: A Search Algorithm for Propositional Satisfiability. IEEE Transactions on Computers 48, 506–521 (1999)
10. Tessaris, S.: Questions and Answers: Reasoning and Querying in Description Logic. PhD Thesis, Department of Computer Science, The University of Manchester, UK (2001)

Identifying Minimal Genomes and Essential Genes in Metabolic Model Using Flux Balance Analysis

Abdul Hakim Mohamed Salleh[1], Mohd Saberi Mohamad[1],
Safaai Deris[1], and Rosli Md. Illias[2]

[1] Artificial Intelligence and Bioinformatics Research Group, Faculty of Computer Science and Information Systems, Universiti Teknologi Malaysia, Skudai, 81310 Johor, Malaysia
ahakim26@live.utm.my, {saberi,safaai}@utm.my
[2] Department of Bioprocess Engineering, Faculty of Chemical Engineering, Universiti Teknologi Malaysia, 81310 UTM Skudai, Johor, Malaysia
r-rosli@utm.my

Abstract. With the advancement in metabolic engineering technologies, reconstruction the genome of a host organism to achieve desired phenotypes for example, to optimize the production of metabolites can be made. However, due to the complexity and size of the genome scale metabolic network, significant components tend to be invisible. This research utilizes Flux Balance Analysis (FBA) to search the essential genes and obtain minimal functional genome. Different from traditional approaches, we identify essential genes by using single gene deletions and then we identify the significant pathway for the metabolite production using gene expression data. The experiment is conducted using genome scale metabolic model of Saccharomyces Cerevisiae for L-phenylalanine production. The result has shown the reliability of this approach to find essential genes for metabolites productions, reduce genome size and identify production pathway that can further optimize the production yield and can be applied in solving other genetic engineering problems.

Keywords: Metabolic engineering, minimal genome, essential genes, flux balance analysis, metabolites productions.

1 Introduction

Systems metabolic engineering has been recognized as a new paradigm for systematically designing novel strategies for improvement of microbial strain. This system level of understanding can be used to help researcher prioritize experimental projects to ensure efficiency in cost and time consumed. In silico metabolic engineering has enabled us to generate hypotheses and predictions systematically to ensure laboratory experiment can be conducted with prior knowledge for optimal results. The application of ‘omics’ data for metabolic analysis along with validation to experimental data can be used to evaluate the significance of the model [1].

Many approaches for optimizing microbial strains have been conducted using genome scale metabolic model of an organism. However these approaches did not

A. Selamat et al. (Eds.): ACIIDS 2013, Part I, LNAI 7802, pp. 414–423, 2013.

utilize gene expression analysis to aid in their prediction. Numerous amount of research incorporate the genetic factors that contribute to the function of metabolic networks as proposed by Karp et al. [2] and Mlecnik et al. [3], but they can only identify groups of specified genes are important although only some genes within this known groups are contributing to the observe response. Probabilistic network models such as Markov Random Field [4] and Mixture Model on Graph [5] on the other hand able to confirm that the features to be logically connected within the metabolic network but an assumption has to be made that is the gene expression is discretely distributed. This may not correctly describe the underlying structure and mechanisms of the system.

In this research, we took vanillin production in S. cerevisiae as a case study to test our approach. S. cerevisiae is considered one of the backbones in metabolic engineering as it is widely used in many applications [6]. High worldwide consumptions of vanilla and laborious and time consuming process of harvesting the product has urge the researchers to find a better alternative of microbial host.

The next section of this paper will discuss about the methodology of this research which covers the processes involved in the approach and dataset used. Then it will be followed by experimental results obtained and discussions and finally conclusions which conclude the findings of this research.

2 Methodology

In this paper, we reduced the size of genomes by implementing gene deletion strategies which is not done by previous methods by assuming that smaller number of essential genes in genomes decreased the used of biochemical resources to produce metabolites thus a higher production of metabolites can be yield. In this research S. cerevisiae genome scale model (yeast.4.05.xml) [7] which consists of 1865 reactions and 1319 metabolites is used to show the enhancement of vanillin production.

Here we have chosen vanillin production to test our approach. The process of formation of vanillin is known as biotransformation of aromatic acids. The basic substrate that can be used to produce vanillin in S. cerevisiae is L-phenylalanine thus, we can say that the production of vanillin increased when the production of L-phenylalanine increased [8].

In the genome of an organism, essential genes are genes compulsory to be present and cannot be knockout as they would results in lower growth rates or exactly zero growth rates. One way to determine these genes is by conducting single gene knockout in Yeast metabolic model and determine based on the resulting growth rates of each knockouts. A series of single gene deletions were performed using the model in order to determine the essential metabolic genes. Minimize the genome size with the assumption that smaller genome size will have less competing production to vanillin.

However, results derived from single gene knockout analysis would not sufficient for us to determine the effect of gene deletion in whole genome. This is because the numbers of genes are large and to perform combinations of multiple gene knockout would take a considerably huge amount of time to complete. By assuming that the genome consist of essential and non-essential genes for a particular process, we can deduce that the genome remain functional as long as the essential genes still exist and those that are not can be neglected or taken out of the system.

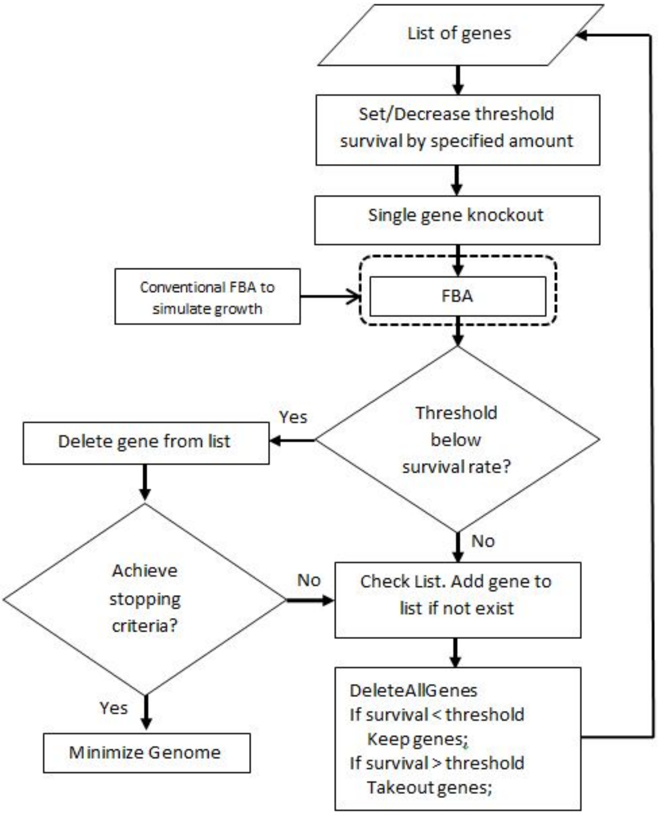

List of genes
Set/Decrease threshold survival by specified amount
Single gene knockout
Conventional FBA to simulate growth
FBA
Threshold below survival rate?
Yes
Delete gene from list
No
Achieve stopping criteria?
No
Check List. Add gene to list if not exist
Yes
Minimize Genome
DeleteAllGenes
If survival < threshold
Keep genes;
If survival > threshold
Takeout genes;

Fig. 1. General framework to minimize model

One of the strategies to address this is to reduce the size of genomes by finding the minimal number of components without reducing the significant functional capabilities. However, this reduction strategy is affected by the order of the genes that have been deleted. For instance, if a particular gene is deleted, it may have caused few other genes that are initially considered unessential to become essential and vice-versa. Another problem to address is the final growth rate. The resulting minimal genome will depend on percentage of final growth rate required, the larger the percentage, the size of minimal genome can afford to be quite large. Figure 1 shows the framework to obtain essential genes and minimal genomes.

It starts with a random model gene by deleting the gene and assessing the resulting growth rate. If the growth rate is considered acceptable (survival rate greater than threshold), the gene is permanently be deleted or otherwise it is placed back into the model. Then, new random gene is selected. When the resulting growth rate is calculated using FBA (dashed box in Figure 1), it is not only assessing the impact of this one gene deletion. It assesses the impact of all previously permanently deleted genes. The same process is repeated until all genes have been accounted. Next, the process is repeated again but with a lower threshold. The deletion cycle continue until the final growth rate is reached.

Next, we utilize KEGG database as our main reference for our pathways that is going to be extracted using microarray gene expression data. This experiment is based on the framework proposed by Hancock et al. [9] where more detailed explanation can be seen. In the initialization phase, the pathway structure is define where each gene is defined as node in the network and annotated by its gene code (*G*), reaction (*R*) and KEGG pathway membership (*P*). On the other hand, the edges that connect the nodes are identified as first substrate compound (C_F); the product compound of first reaction (C_M); final product compound (C_T) and the final KEGG pathway membership of C_T, (*P*) as in Eq. (1).

$$nodes = (G, R, P);\ edges = (C_F, C_M, C_T, P) \qquad (1)$$

In Eq. (2), the probability of *y*, a binary response variable given that *X*, which is a binary matrix where the columns represent genes, the rows represent a pathway, and value of one indicates that the particular gene is included within specific path is defined that consist of two parts. First is the sum of probability π_m, which is the probability of each component with *y* given that *X* with βm parameter and second, product of $p(g_k,\ label_k|g_{k-1};\ \theta_{km})$, which is the probability of path travers on edge $label_k$. g_k denotes the current gene and next gene in sequence, g_{k+1} where $label_k$ is the edge annotation. The result of this 3M (Markov Mixture Model) is *M* components defined by $\theta m = \{\theta_{sm}, [\theta_{2m},\ldots, \theta_{tm},\ldots, \theta_{Tm}]\}$. The θ_m is probabilities of each gene clustered within each component and indicate the importance of the genes. The parameters π_m, θ_{km} and βm are estimated simultaneously with an EM algorithm where more detailed explanation are discussed by Hancock and Mamitsuka [10].

$$p(y \mid X) = \sum_{m=1}^{m} \pi_m p(y \mid X, \beta_m) \prod_{k=2}^{k} p(g_k, label_k \mid g_{k-1}; \theta_{km}) \qquad (2)$$

By using set of genes that involved in the particular pathway, *p-values* for each pathway are calculated using the hypergeometric distribution by summation of binomial coefficient. If the whole genome has a total of (*m*) genes, of which (*t*) are involved in the pathway under investigation, and the set of genes submitted for analysis has a total of (*n*) genes, of which (*r*) are involved in the same pathway, (*x*) is the number of pathway that have been chosen. Then the *p-value* can be calculated to evaluate enrichment significance for that pathway by Eq. 3:

$$p = 1 - \sum_{x=0}^{r-1} \frac{\binom{t}{x}\binom{m-t}{n-x}}{\binom{m}{n}} \qquad (3)$$

FBA is a widely used and basic approach of constraints-based flux analyses and have shown to be successful for predicting growth, uptake rates and by-product secretion [11] without requiring the knowledge of metabolite concentration or the enzyme kinetics details of the system. FBA uses Linear Programming (LP) to maximize an objective function under different constraints. For example, to optimize an objective function denotes by *Z* at a particular period of time, *c* and *v* is reaction involved (*e.g* growth) typically the LP is formalized as in Eq. (4) and (5):

$$Z = \sum_{i=1}^{r} c_i v_i \tag{4}$$

$$S \,.\, v = 0,\ v_{min} \leq v \ \leq v_{max}. \tag{5}$$

Mass balance constraints are imposed by a system of linear equation, where stoichiometric *S* is an *m x n* matrix where *m* is the number of metabolites, and *n* is the number of reactions. *vmin* and *vmax* are set as lower and upper bounds on flux values that impose thermodynamic constraints that restrict directional flow of reaction, and capacity constraints. Using the minimal genome that consists of essential genes earlier and also the pathway membership, we initialized the stoichiometry matrix based on both of the results obtained. Using flux balance analysis we then calculate the optimization for our objectives function that is L-phenylalanine production and growth rate.

3 Experimental Results and Discussion

The experiment is conducted using glucose minimal media. The results obtained shows that the minimal genomes has an average size of 300 genes, approximately 30% of the original size. After running the experiment with threshold of original growth rates, there are about 130 genes of minimal genome are detected as essential genes out of the original 924 genes that produce growth rates of 1.3276 mmol gDW^{-1} hr^{-1}. Figure 2 shows the number of genes for 10 runs compared to the original number for L-phenylalanine production.

Logically, since only single gene deletion is performed we could not entirely conclude that the genome will survive and operates with only these genes because the number is considerably small compared to the original number of genes. Furthermore, the knockout process is dependent on just a single gene and the sequence of the genes. For example, if single gene A or B is knocked out, the cell may still survive but what if both of them are knocked out. Hence, the numbers of combinations are big.

Therefore, the size and contents of the minimal genomes might be varied depending on which genes are deleted first. Another factor to consider is the final growth rate. If the desired growth rate is 90% of it was originally, the resulting minimal genome may be quite large. If much smaller growth rate is desired, then the minimal genome can afford to become much smaller. Another factor is the growth medium. Highly enriched media will aid in achieving a smaller minimal genome.

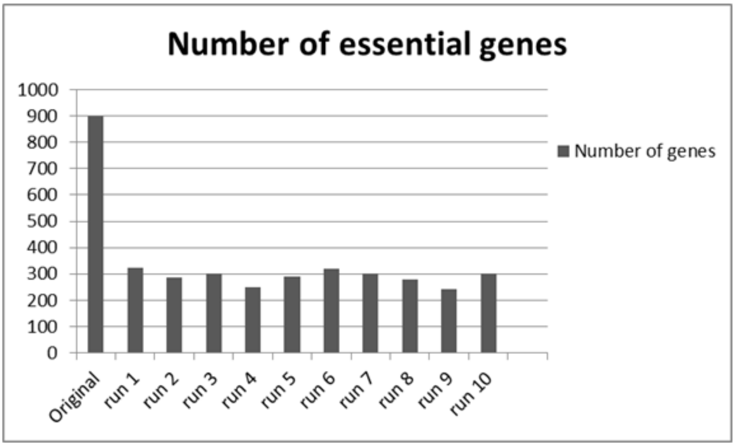


Fig. 2. Number of genes in minimize genomes

Existence of a large number of flux routes or pathways in genome-scale metabolic models requires the use of optimization or computational methods to predict the alternative routes consist of essential genes and deletion of genes in which help to improve the production. FBA is guaranteed to produce optimal results but not necessarily unique due to the existence of a large number of pathways involves [12]. With the essential genes obtained we extract significant pathways that lead to the production of L-phenylalanine using microarray gene expression data. Figure 3 shows the pathway extracted that consist of compound names as the nodes and KEGG reaction numbers for L-phenylalanine.

It is clear from set of compounds that made up the pathway, the highest path probability would be the transition and conversion alpha-D-glucose as the source moving towards the whole glycolysis pathway to produce pyruvate, CO_2 and Urea. Pyruvate plays a critical role in balancing between fermentation and respiration and also a potential intracellular indicator for limitation of glucose [13]. Then NH_3 is formed which leads to production of L-tyrosine and L-phenylalanine as final products.

Using set of genes that involve in the pathway we calculate the p-value for the whole genome to investigate furthermore which metabolism are actually contributing to the metabolite production. The *p-value* obtained can be used to measure the gene membership in the pathway. Table 1 shows the top 5 pathways correspond to that particular set of genes.

From the table it is obvious that Phenylalanine metabolism pathway has the lowest *p-value* with the highest gene ratio indicating the significant of the pathway with the gene set produce by the experiment. The pathways are considered to be highly statistically significant if having $p < 0.01$. This observation is probably caused by the production of L-phenylalanine and vanillin itself is a part of the component of the metabolism system therefore more number of genes is detected within this particular metabolism.

Figure 4 shows the result of glucose uptake rate effect towards the growth rate for both, new model (solid line) and original (dashed line). At the initial stage, with glucose uptake of 0 mmol gDW^{-1} hr^{-1} the maximum possible growth rate is 0 hr^{-1}. At approximately 18 to 20 mmol gDW^{-1} hr^{-1} which is the biologically realistic uptake rate [14] we can see the the production of L- phenylalanine is slightly higher with 1.3276 mmol gDW^{-1} hr^{-1} compared to the original 1.1596 mmol gDW^{-1} hr^{-1}.

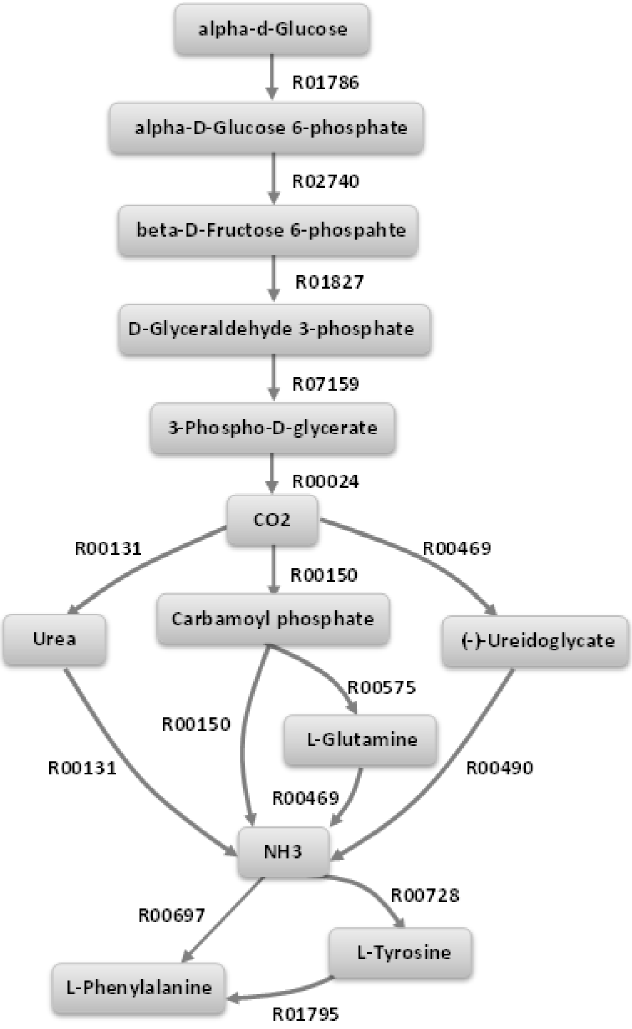


Fig. 3. Significant metabolic pathway for L-phenylalanine production based on KEGG

Table 1. The pathway membership for L-phenylalanine based on KEGG pathways

PATH	PATHWAY NAME	GENE RATIO	BACKGROUND RATIO	P-VALUE
00360	Phenylalanine metabolism	41/221	161/4377	1.11E-16
00010	Glycolysis / Gluconeogenesis	35/221	79/4377	4.44E-16
00020	TCA Cycle	21/221	58/4377	6.22E-15
00250	Alanine, aspartate and glutamate metabolism	14/221	35/4377	1.63E-14
00062	Arginine and proline metabolism	13/221	29/4377	1.89E-14

Normally, the biochemical production would increase along with cellular growth rate [15] hence, this indicate that the model able to survive and produce the desired products at optimal rate. Then the growth start to increase rapidly when enough glucose is available in the system meaning that the amount of ATP produce for growth has meet its requirement. After a certain period, the growth starts to increase less rapidly at one point until the end. This is due to the fact that at that particular point glucose is no longer the limiting factor for growth but instead its oxygen. In this condition the access glucose produce cannot be fully oxidize thus changing the flux to the production pathways.

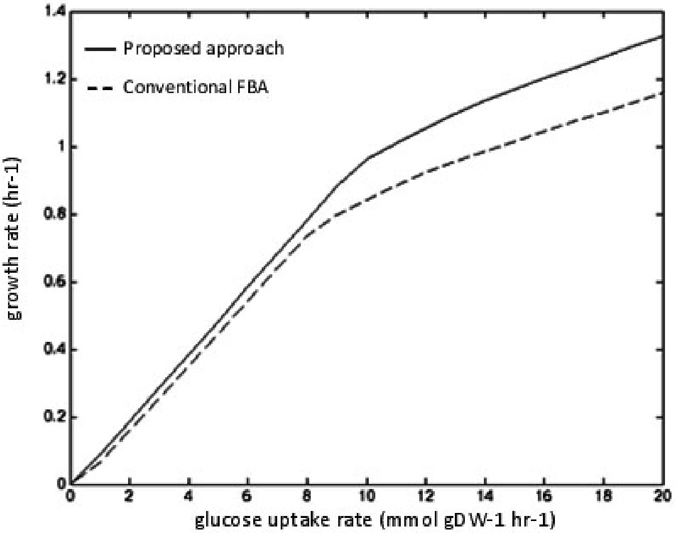


Fig. 4. Glucose uptake rate effect towards the growth rate for L-phenylalanine production

4 Conclusion

In this paper, we proposed an approach to identify the essential genes that able to form a minimal genome without degrading the biological function using FBA. Then, based on the essential genes obtained, we construct a metabolic pathway from gene expression data for a particular production of metabolites of interest. FBA is used to produce a fitness function with the assumption that the genome is in a steady state condition whereby optimization of the objective functions, in this case L-phenylalanine production can be conducted.

Based on the experiment conducted on S. cerevisiae for L-phenylalanine production, the results shown that the information provided by gene expression analysis has improve the prediction of constraint based analysis such as FBA and can potentially be extend. The integration of different data such as gene expression data, transcriptional regulatory and metabolic flux data has also shown to be successful in metabolic engineering for various purposes. Hence, the next big challenge would be integrating these models to a more biologically significant representation of these interrelated networks.

Acknowledgments. We would like to thank Universiti Teknologi Malaysia for supporting this research by UTM GUP research grants (vot numbers: QJ130000.7123.00H67 and QJ130000.7107.01H29).

References

1. Edward, J.S., Ibarra, R.U., Palsson, B.O.: In silico predictions of Escherichia coli metabolic capabilities are consistent with experimental data. Nature Biotechnology 19, 125–130 (2001)
2. Karp, P.D., Paley, S.M., Krummenacker, M., Latendresse, M., Dale, J.M., Lee, T.J., Kaipa, P., Gilham, F., Spaulding, A., Popescu, L., Altman, T., Paulsen, I., Keseler, I.M., Caspi, R.: Pathway tools version 13.0: integrated software for pathway/genome informatics and systems biology. Brief Bioinform. 11(1), 40–79 (2010)
3. Mlecnik, B., Scheideler, M., Hackl, H., Hartler, J., Sanchez-Cabo, F., Trajanoski, Z.: PathwayExplorer: web service for visualizing high-throughput expression data on biological pathways. Nucleic Acids Research 33(1), 633–637 (2005)
4. Wei, Z., Li, H.: A markov random field model for network-based analysis of genomic data. Bioinformatics 23(12), 1537–1544 (2007)
5. Sanguinetti, G., Noirel, J., Wright, P.C.: Mmg: a probabilistic tool to identify submodules of metabolic pathways. Bioinformatics 24(8), 1078–1084 (2008)
6. Varges, F.A., Pizzarro, F., Perez-Correa, J.R., Agosin, E.: Expanding a dynamic flux balance model of yeast fermentaion to genome-scale. BMC Systems Biology 5, 75 (2011)
7. Mo, M.L., Palsson, B.Ø., Herrgård, M.J.: Connecting extracellular metabolomic measurements to intracellular flux states in yeast. BMC Systems Biology 3, 37–41 (2009)
8. Priefert, H., Rabenhorst, J., Steinbüchel, A.: Biotechnological production of vanillin. Appl. Microbiol. Biotechnol. 6, 296–314 (2001)
9. Hancock, T., Takigawa, I., Mamitsuka, H.: Mining metabolic pathways through gene expression. Gene Expression 26(17), 2128–2135 (2010)

10. Hancock, T., Mamitsuka, H.: A Markov Classification Model for Metabolic Pathways. In: Salzberg, S.L., Warnow, T. (eds.) WABI 2009. LNCS, vol. 5724, pp. 121–132. Springer, Heidelberg (2009)
11. Reed, J.L., Palsson, B.O.: Thirteen Years of Building Constraint-Based InSilico Models of Escherichia coli. J. Bacteriol. 185(9), 2692–2699 (2003)
12. Brochado, A.R., Matos, C., Moller, B.L., Hansen, J., Mortensen, U.H., Patil, K.R.: Improved vanillin production in baker's yeast through in silico design. Microbial Cell Factories 9, 84 (2010)
13. Boer, V.M., Crutchfield, C.A., Bradley, P.H., Botstein, D., Rabinowitz, J.D.: Growth-limiting intracellular metabolites in yeast growing under diverse nutrient limitations. Mol. Biol. Cell 21(1), 198–211 (2010)
14. Orth, J.D., Thiele, I., Palsson, B.Ø.: What is flux balance analysis? Nature Computational Biology 28, 245–248 (2010)
15. Kim, J., Reed, J.: OptORF: Optimal metabolic and regulatory perturbations for metabolic engineering of microbial strains. BMC Bioinformatics 4(53), 1–19 (2010)

A New Data Hiding Scheme for Small Blocks of Twelve Pixels on Binary Images by Module Approach

Phan Trung Huy[1], Nguyen Hai Thanh[2], Le Quang Hoa[1], and Do Van Tuan[3]

[1] Hanoi University of Science and Technology, Hanoi - Vietnam
huyfr2002@yahoo.com, huypt-fami@mail.hut.edu.vn
[2] Ministry of Education and Training
nhthanh@moet.gov.vn
[3] Hanoi College of Commerce and Tourism, Hanoi - Vietnam
dvtuanest@gmail.com

Abstract. Color and binary images are popular subjects for data hiding. In this paper we propose a method to controlling quality of hiding secret data in binary images by using small blocks of 12 pixels with a restriction that in each block at most two pixels are changed to embed 6 secret bits in case it is successful. The method uses the $\mathbf{Z}_2$-module $\mathbf{Z}_2^6$ of integers modulo 2 and this shows the way to overcome the limit (12,6)2 which can not reached by error correcting block codes approach. The method presents the way used secret keys for small blocks which remain high safety property of the scheme. The scheme can be modified to control high quality of stego palette images by combining with Optimal Parity Assignment (OPA) approach for hiding secret data in palette images, to prevent from steganalysis.

Keywords: data hiding, binary image, high ratio, controlling quality, module approach, 2-weak base.

1 Introduction

In security area, data hiding gives us a way to provide high safety levels for keeping secret data invisibly when they are transmitted in internet, networks. In applications, digital image is a popular format to hide data in. The most challenging problem is to establishing a hiding scheme which allows us to hide secret data into images with a high ratio of secret data and a low distortion of stego-images as much as possible.

In a data hiding scheme based on block-based approach, for example see[3], ones can split the data area of a given image into several separated blocks with the same size, and try to hide as much as secret data in. It is better if as large amount of data embedded in each block as possible with a very small number of changed pixels that means the PSNR value of the method need to large enough. In binary images, embedding secret data is very difficult for protection from revealing by human eyes.

Once there exits a hiding scheme in binary scheme, ones can modify it to build a modified hiding scheme for palette images with a high quality, which uses matrices of bits generated from these images by using Parity Approach (PA) [4] and Optimal

A. Selamat et al. (Eds.): ACIIDS 2013, Part I, LNAI 7802, pp. 424–434, 2013.

Parity Approach (OPA) [5,7]. Since the number of colors in palettes are often small, to prevent from steganalysis, especially to histogram-based attacks, as shown in [8,9].., it is very difficult to test whether the image contains hidden data or not, if the ratio of pixels changed to the total pixels of each block of the given image is not larger than about 1/5 (the smaller the better).

In this paper, we propose a 2-hiding scheme which is a scheme, in each block of 12 pixels of binary images (*twelve block*), permits at most 2 pixels changed to hide 6 secret bits, in cases it is successful. To control the quality of stego-images, in a block, among 6 embedded bits, 1 bit is remained for controlling the quality of the stego-image, 5 rest bits are real secret bits. The way to control the quality is the same as shown in our previous work [6] which is a modification of the method introduced by Tseng-Pan [10]. Let us remak that the existence of a covering code COV(12,6,2) is an open problem arisen in [1] and there is a closed result in [2] which used interesting algebraic techniques to show that t(12,6)= 3 is the largest error can be corrected by a linear correcting code of type [12,6] (the dimension of the block code is 6 in the linear space of dimension 12). As shown in [5], this means that if we use this code to apply in a hiding scheme, we can only hide 6 bits in a block of 12 pixels by changing at most 3 pixels, this rate can not be reached in the case if we permit only at most two pixels changed. But this is the claim we need in this paper to build 2-hiding schemes, since the less number of pixels changed, the less the distortion of the stego-image introduced, the more safety we obtain for the hiding scheme. In our proposed scheme, the Z_2-module M= $\mathbf{Z}_2^6$ is used by module approach and by introducing a notion of 2-weak base in M, which allows us to present 63 elements among 64 elements of M as combination of at most 2 elements in the base. The results obtained show the optimism of module approach to data hiding area.

For the presentation of the paper, after the introduction section 1, in section 2 we recall 2- hiding schemes for data hiding based on modules over rings of characteristics 2 [6] and the relationship with the result setlled in [2] by A. R. Calderbank and N. J. A. Sloane for correcting codes. In section 3, we introduce the notion of 2-weak base of modules and present applications of these bases, 2-bases and 2-weak bases in hiding schemes by establish two algorithms for hiding data in binary images and the way uses secret keys for small blocks which remains high safety property of the scheme. This shows the fit of 2-weak bases in our hiding scheme with controlling high quality of stego-images and high rate of embedded bits.

In the section 4, some experimental results are presented to show the advantage of our scheme. Together with this scheme, we consider the way to apply Parity Approach (PA) due to our previous work [7] or works in [4,5] to build a new schemes applied for palette images. Discussion about the developing of some future works is mentioned in the conclusion section 5.

2 Application of Z_2-Module in Hiding Data

Recall that each (right) module M over the ring $\mathbf{Z}_q$ is an additive abelian group M with zero 0 together with a scalar multiplication "." to assign each couple (m,k) in $M\times\mathbf{Z}_q$ to an element $m.k$ in M. Let $\mathbf{Z}_q=\{\mathbf{0},\mathbf{1},..,\boldsymbol{q}\text{-}\mathbf{1}\}$. Some following basic properties are used usefully in the sequent:

P1) $m.\mathbf{0} = 0$; $m.\mathbf{1}=m$;

P2) $m+n = n+m$ for all m,n in M.

P3) $m.(\boldsymbol{k}+\boldsymbol{l}) = m.\boldsymbol{k} + m.\boldsymbol{l}$ for all m in M, $\boldsymbol{k},\boldsymbol{l}$ in $\mathbf{Z}_q$.

Definition 2.1. Let U be a subset of M, $0 \notin \mathrm{U}$. We call U

i) A *2-base* of M if any element in M can be presented as a linear combination of as most two elements in U.

ii) A *2–weak base* of M if almost elements in M can be presented as a linear combination of as most two elements in U. That is there is a subset V of M, $|M\text{-V}|$ is small, such that any element in V can be presented as a linear combination of as most two elements in U.

Given an image G, denote by C_G the set of colors of G, $C_G= \{C_p: p \text{ in } G\}$ where each C_p is the color of pixel p. Suppose that we can find a function Val: $\mathrm{C}_G \to \mathbf{Z}_q$ and a *color changing mapping* Next: $\mathrm{C}_G \to \mathrm{C}_G$ satisfying the condition:

(2.1) $\forall c \in \mathrm{C}_G$, Val (Next($c$)) =Val($c$)+1.

And for the palette image case we claim one extra condition:

(2.2) $\forall c \in \mathrm{C}_G$, c' =Next(c) is a color “similar” to c.

For binary images, applying q=2 the addition in $\mathbf{Z}_2$ can be seen as the operation exclusive -OR on bits, and $M=\mathbf{Z}_2 \times \mathbf{Z}_2\times..\times \mathbf{Z}_2$ is the n-fold cartesian product of $\mathbf{Z}_2$ which can be seen as a (right) $\mathbf{Z}_2$-module, each element $x=(x_1,x_2,..,x_n)$ in M can be presented as an n-bit stream $x=x_1x_2..x_n$, with operations defined by:

D1) For any $x=x_1x_2..x_n$, $y=y_1..y_n$ in M, $\boldsymbol{k}$ in $\mathbf{Z}_2$, $x+y = z_1z_2..z_n$ where $z_i=x_i +y_i$, i=1,..,n can be computed by bitwise XOR.

D2) $x.\boldsymbol{k}= z_1z_2..z_n$ where $z_i=\mathrm{x}_i.\boldsymbol{k}$ (= x_i AND $\boldsymbol{k}$).

Given a binary image G, we set $C_G= \mathbf{Z}_2 =\{\mathbf{0},\mathbf{1}\}$ and Val is the identical function on $\mathbf{Z}_2$, Val(c)=c for all c in $\mathbf{Z}_2$. The function Next: $\mathbf{Z}_2 \to \mathbf{Z}_2$ is defined by

(2.3) Next(c)=c+1, (also $c\oplus 1$), for all c in $\mathbf{Z}_2$

and *changing a color c* is done by replacing c with c'=Next(c)=c+1.

Given a set $S=\{p_1,p_2,..,p_N\}$ of N pixels in a binary image G, a *2-base* U of $M = \mathbf{Z}_2 \times \mathbf{Z}_2\times..\times \mathbf{Z}_2$-the n-fold cartesian product of $\mathbf{Z}_2$ as above, $N \geq |U|$, we define a *weight function* as a surjective mapping:

(2.4) h: $\{1,2,..,N\} \to U$ for which, each p_i in S, $m=h(i)$ is called *the weight of* p_i.

Given a set $S=\{p_1,p_2,..,p_N\}$ of N pixels in a binary image G, a *2-base* U of $M = \mathbf{Z}_2 \times \mathbf{Z}_2\times..\times \mathbf{Z}_2$-the n-fold cartesian product of $\mathbf{Z}_2$ as above, a secret set $K=\{k_i \in \mathbf{Z}_2\text{: } 1\leq i \leq N\}$ of N key bits k_i , and a weight function h: $\{1,2,..,N\} \to U$. We can hide a n-secret bit stream $b= b_1b_2..b_n$ by changing color of at most two pixel in S in the following 2-hiding scheme.

2.1 Hiding the Secret Item b into S

Step 0) Change the color C of each pixel $p_i \in S$ into a new color $C_i^*=C_i+k_i$ (in $\mathbf{Z}_2$). We present this result as T=S ⊕K;

For the new colors of pixels in T:

Step 1) Computing $m = \sum_{1\leq i \leq N} h(i).C_i^*$ in the $\mathbf{Z}_2$ –module M.

Step 2) Case $m = b$: keep S intact;

Case $m \neq b$: Since U is a 2-base of M, two cases can be happened:

a) There is $p_i \in S$ such that $h(i)= b\text{-}m$:
We change the color C_i of p_i into $C_i^{'}$ = Next(C_i)=C_i+1; { only 1 pixel p in S is changed, after hiding b into S , then the new color of p_i is $C_i^* = C_i'+k_i = C_i+1+k_i=C_i+k_i+1=C_{px}^*+1$}

b) There is $p_{i,}p_j \in S$ such that $h(i)+h(j)= b\text{-}m$:
We change the color C_i of p_i into $C_i^{'}$ = Next(C_i)=C_i+1, C_j of p_j into $C_j^{'}$ = Next(C_j)=C_j+1; {2 pixel p in S are changed, after hiding b into S}

2.2 Extracting the Secret Item d from *S*

Step 0) Using the secret set K, change the color C_i of each $p_i \in S$ into a new color $C_i^*=C_i+k_i$. For the new colors of pixels in T, apply steps 1),2) in **2.1.** as follows:

Step 1) Computing $u = \sum_{1\leq i \leq N} h(i).C_i^*$ in the $\mathbf{Z}_2$ –module M.

Step 2) Return $b=u$.

Correctness of the Method

Since h is surjective and U is a 2-base, as in [6] we deduce

Proposition 2.1. The secret item d embedded in the stego-block S in the phase 2.1 can be extracted by using the phase 2.2 above.

2.3 Hiding Data in a Sequence of Binary Blocks

For the case each binary block of the image G has a small size N, we can use a sets of binary keys of a large size $p.N$ to apply in real applications, preventing from revealing keys by exhaustive attacks. Indeed, suppose the image G is spitted into separated blocks of N pixels, written as a sequence $B_1,B_2,..,B_k$. We use a secret set K of $p.N$ bits as a common key set and split it into p subsets $K_0,..,K_{p\text{-}1}$ of the same size N.

For any sequence of n-bit strings $s_1,s_2,..,s_q$, $q \leq k$, we can apply step by step the algorithms above for hiding and extracting embedded data as follows:

Data Hiding Algorithm A1:

```
Step 1) $t$ =0;
Step 2) For $i$=1 to q do
    a) Hide the n-bit string $s_i$ in $B_i$ by using the binary subset key $K_t$ by the phase
       2.1. above;{some first $s_i$'s are supposed to use for saving the size $q$ and extra
       information}
    b) $t = (t+1)$ mod p;
```

Data Extracting Algorithm A2:

```
Step 1) $t = 0$; $i$=1;
Step 2) While can estimate $i \le q$ do
        a) Extract the n-bit string $s_i$ in $B_i$ by using the binary subset key $K_t$ by the
        phase 2.1. above;
        b) Use some first n-bit string $s_i$'s for determining the size $q$ and other
        information if one need;
        c) $t = (t+1) \bmod p$;
        d) $i=i+1$;
        Endwhile;
```

2.4 Correcting Code Technique to 2-Hiding Schemes

Recall that a binary, linear block code C with block length n and dimension k is denoted as an $[n; k]$ *code*, and one call *an* $[n; k; d]$ *code* if its minimum distance is d. The code's *covering radius* $r(C)$ can be defined as the smallest number r such that any binary column vector of length $(n - k)$ can be written as a sum of r or fewer columns h_i, i=1,..,n, of a *parity check matrix H* of C, $H=[h_1,h_2,..,h_n]$. An $[n; k]$ code with *covering radius r* is denoted an $[n; k]r$ *code*. In [2], $l(m; r)$ is defined to be the smallest n such that an $[n; n\text{-}m]r$ code exists. We can use an $[n; k]r$ *code* with its *parity check matrix H* of C, $H=[h_1,h_2,..,h_n]$, to an r-hiding scheme which permit us to hide $n\text{-}k$ bits in each block of n pixels of a binary image by changing at most r pixels, as follows.

Hiding Phase:

($n\text{-}k$ bit string x is embedded in a block u of n pixel)

Given an $n\text{-}k$ bit string x as secret data which needs to be embedded in a binary block u of n pixels.

a) Present x as a binary column vector of dimension $n\text{-}k$ and u as a column vector $u = (u_1,u_2,..,u_n)^t$ of dimension n over the field $\mathbf{Z}_2$.

b) Compute the syndrome $m = H.u = h_1.u_1+..+h_n.u_n$ of u, m as a sum in $\mathbf{Z}_2$-module $M = \mathbf{Z}_2 \times \mathbf{Z}_2 \times .. \times \mathbf{Z}_2$-the $n\text{-}k$-fold cartesian product $\mathbf{Z}_2^n$ of $\mathbf{Z}_2$, where $h_1,..,h_n \in M$.

Two cases happen:

b1) $m = x$. Then u is kept intact, considered as the stego-block in which x is hidden.

b2) $m \neq x$.

- compute $y = x - m$ in M;
- Finding an error column vector e of *hamming weight* at most r such that $H.e = y$, by correcting code technique.
- Put $u' = u + e$ (or, by changing at most r binary pixels, those are r positions in u), we have u' as the stego-block in which x is hidden.

(Let us note that $H.u'=H(u+e)=H.u+H.e=m+y = x$ as the definition of y above)

Extracting Phase

a) Given the stego-block u', present u' as above.

b) Extract $m = H.u'$ as the secret $n\text{-}k$ bit string embedded in u' (that is $m = x$).

Remak 2.1. 1) The two phases above are nothing but the special cases of two phases 2.1, 2.2 without considering extra binary keys. Hence we can add a secret binary key set of *n* elements for real applications. This example shows the closed relationship between the two approachs, where module approach can be considered as a general case of correcting code approach.

2) The results that $l(6,2) = 13$ and $t(12,6)=3$ in [2] give us a conclusion: it is impossible to find a [12;6]2 code to apply in data hiding, for a 2-hiding scheme in which we use each block of 12 pixels to hide 6 bits by changing at most 2 pixels.

3) Closely with this result, in [5], Fridrich gave an open problem: does there exist a COV (2, 12, 6) which allows us to hide 6 bits in each binary block of 12 pixels by changing at most two pixels?

In the next section, we introduce a new 2-hiding scheme with controlling quality which allows us to hide 6 bits in each binary block of 12 pixels by changing at most two pixels, where one embed bit is used as a flag bit which informs us whether the block is successful or not for hiding data. In case successful, 5 rest bits are used for real applications. This scheme uses a 2-weak base of the module $M = \mathbf{Z}_2^6$, but we can see that the total number of bits hidden in a binary image G in almost binary blocks approximates the limit of COV(12,6,2) if this COV exits.

3 Application of 2-Weak Bases of Z₂-Modules for Hiding Schemes

In this section, we introduce a method using 2 -weak bases in hiding scheme with controlling quality of binary stego-images.

For the module $M= \mathbf{Z}_2^6$ over the field $\mathbf{Z}_2$(its elements can be seen as natural numbers or 6-binary strings or column vectors of dimension 6. The following 2-weak base of 12 elements is obtained by our program:

$U = \{11, 51, 55, 39, 30, 29, 1, 2, 4, 8, 16, 32\}$ in natural numbers, or

$U =\{w_1, w_2, w_3, w_4, w_5, w_6, w_7, w_8, w_9, w_{10}, w_{11}, w_{12}\}$, w_1=001011, w_2=110011, w_3=110111, w_4=100111, w_5=011110, w_6=011101, w_7=000001, w_8=000010, w_9=000100, w_{10}=001000, w_{11}= 010000, w_{12}= 100000, in binary representation.

With this base, except the element 52 with binary presentation as 110100, all others elements in *M* can be presented as a linear combination of at most 2 elements of *U*. This property help us in almost binary blocks of 12 pixels we can hide 6 bits and we can use this to build a 2-hiding scheme with controlling quality of stego-image as well.

3.1 Parameters of the Scheme

We present the 2-weak base *U* as the weight matrix of the scheme, a binary block *S* of 12 pixels as a binary matrix where we need to hide a 6-bit string *D* in, a binary matrix key *K*, each matrix has the same size 4 × 3, k_i and p_i, i=1,..,12, can be considered as elements in $\mathbf{Z}_2$. For brevity, the natural numbers can be used instead of binary representation of elements in *U*. In *U*, all odd elements are underlined.

$$U = \begin{array}{|c|c|c|}\hline \underline{w_1} & w_8 & \underline{w_6} \\ \hline w_9 & \underline{w_2} & w_{10} \\ \hline \underline{w_4} & w_{11} & \underline{w_3} \\ \hline \underline{w_7} & w_5 & w_{12} \\ \hline \end{array} \qquad S = \begin{array}{|c|c|c|}\hline p_1 & p_8 & p_6 \\ \hline p_9 & p_2 & p_{10} \\ \hline p_4 & p_{11} & p_3 \\ \hline p_7 & p_5 & p_{12} \\ \hline \end{array} \qquad K = \begin{array}{|c|c|c|}\hline \underline{k_1} & k_8 & \underline{k_6} \\ \hline k_9 & \underline{k_2} & k_{10} \\ \hline \underline{k_4} & k_{11} & \underline{k_3} \\ \hline \underline{k_7} & k_5 & k_{12} \\ \hline \end{array}$$

The same idea in the method given by Tseng-Pan in [10], suppose that we need to hide $D_0 = b_5b_4b_3b_2b_1$ a secret 5-bit string in S. We can add one more control bit $b_0= 1$ as a LSB bit of $D = b_5b_4b_3b_2b_1b_0$ and we try to hide D in S by the same phase 2.1. If it is not success, we flip b_0 to 0 and try to hide any 6- bit string having $b_0=0$ as its LSB in S, to mark that S is a failure block and do not to use the extracted data.

The following conditions and facts are needed for correctness and success of the method.

Condition 1. (odd condition- as in [10]) In each 2×2 submatrix of U, there is an odd entry (with odd value).

Condition 2. (even condition) Corresponding to the 6 (even) number of odd values in U , K has six elements k_1, k_2, k_3, k_4, k_6, k_7 satisfied the sum $k_1+ k_2+ k_3+k_4+ k_6+k_7$ is an even value.

Fact 1. With the condition 2, in case S has all entries are 0 or are 1 (mono-value), the sum [S,K] = $\sum_{1\leq i\leq 12} w_i.(p_i \oplus k_i)$ always even.

Fact 2. Except the element $52=110100_2$, any element in M can be presented as a linear combination of at most 2 elements of U.

Fact 3. In case S is not mono-value, by conditions 1,2, one can always find successfully a p_i among 6 entries p_1, p_2, p_3, p_4, p_6, p_7 in S which closes to another $p_j \neq p_i$ both belong some 2 ×2 submatrix of S.

Now, the hiding and extract phases of the new hiding scheme can be presented as follows.

3.2 Data Hiding Algorithm B1

```
Step 1) Prepare $D = D_0 \times 2+1= b_5b_4b_3b_2b_1 1$ in binary presentation as an element in $M$;
Step 2) Compute the binary matrix of the same size 4 x 3:
        $T = S \oplus K$; {$T$ has entries $t_1,..,t_{12}$ in the same order as positions in $S$ and $K$}.
Step 3) {Try to hide $D$ in $S$} Compute the sum in $M$:
        $x = \sum_{1\leq i\leq 12} w_i.t_i;$
    3.1) Case $x = D$: keep $S$ intact; {$S$ is considered as success block to hide $D0$ in}
    3.2) Case $x \neq D$: Then $y=D \oplus x \neq 0$; {by bitwise XOR in $M$}
      3.2.1) Find $w_i$ in $U$ such that $w_i = y$:
          - If there is such a $w_i$ so that $p_i$ closes to another $p_j \neq p_i$ , they are in the
          same a 2x2 submatrix of $S$, then flip $p_i$ to $1 \oplus p_i$ ;
          - if there is not any such $w_i$ and $p_i$, go to the next;
      3.2.2) Find $w_i$ and $w_j$ in $U$ such that three following conditions hold:
        a) $w_i \oplus w_j = y$;
        b) $p_i$ closes to another $p_j \neq p_i$, they are in the same a 2×2 submatrix of $S$;
        c) $p_k$ closes to another $p_l \neq p_k$, they are in the same a 2×2 submatrix of $S$;
```

```
Then flip two entries, $p_i$ to $1\oplus p_i$ and $p_j$ to $1\oplus p_j$;
 In case we can not find such $w_i$, $w_j$ and $p_i$, $p_j$, go to the next;
3.2.3) Try to mark $S$ as failure:
   a) If x is even: keep $S$ intact and exit;
   b) If x is odd: find one entry $p_i$ in 6 entries $p_1, p_2, p_3, p_4, p_6, p_7$ in $S$ which
     closes to another $p_j \neq p_i$ both in some $2 \times 2$ submatrix of $S$,
     then flip $p_i$ to $1\oplus p_i$ and exit;
```

{In the case $y = D\oplus x = 52$ all subcases 3.2.1 and 3.2.2 above are not successful hence 3.2.3 is chosen}

3.3 Data Extracting Algorithm B2

Given S as a stego-block. We try extracting embedded bits and using it if S is success.

```
Step 1) Compute the binary matrix of the same size 4 x 3:
           $T = S\oplus K$;
Step 2) Compute the sum $x = \sum_{1\leq i\leq 12} w_i.t_i$ in $M$;
   -  If $x$ is odd, return $D_0 = x$ div 2 (remove LSB of $x$) as the hidden 5-bit string
      we need;
   -  If $x$ is even, conclusion $S$ is fail to hide data in and exit;
```

Let us remark that, in the hiding phase 3.2. above if the case 3.2.3 (b) happens, by Fact 2, we always find successfully a p_i in 6 entries p_1, p_2, p_3, p_4, p_6, p_7 in S which closes to another $p_j \neq p_i$ both belong to some 2×2 submatrix of S.

Hence, we can easily deduce the correctness of the scheme by using the conditions 1,2 and Facts 1,2,3 above.

Theorem 3.1. If S is a success block to hide the secret 5-bit string D_0 in by the hiding algorithm B1 in 3.2, ones can always extract successfully D_0 by the extracting algorithm B2 in 3.3.

Remark 3.1. For the most cases, the case $y = D\oplus x = 52$ (in 3.2.3. of the hiding algorithm B1) rarely happens, hence the 2-weak base U can be used successfully as the same way of using COV(12,6,2) if it exits. In any case, we always proceed successfully each steps in two algorithm B1, B2 as claimed.

Example 3.1. Given three matrix as follows

$$U = \begin{array}{|c|c|c|}\hline \underline{w_1} & w_8 & \underline{w_6} \\ \hline w_9 & \underline{w_2} & w_{10} \\ \hline \underline{w_4} & w_{11} & \underline{w_3} \\ \hline \underline{w_7} & w_5 & w_{12} \\ \hline \end{array} \qquad S = \begin{array}{|c|c|c|}\hline \underline{0\ p_1} & 1\ p_8 & 0\ p_6 \\ \hline 1 & \underline{0} & 1 \\ \hline \underline{0} & 0\ p_{11} & \underline{1} \\ \hline \underline{0\ p_7} & 1 & 1 \\ \hline \end{array} \qquad K = \begin{array}{|c|c|c|}\hline \underline{1} & 0 & \underline{0} \\ \hline 1 & \underline{1} & 1 \\ \hline \underline{0} & 0 & \underline{1} \\ \hline \underline{1} & 0 & 0 \\ \hline \end{array}$$

In the hiding phase, T=[$S\oplus K$] has 12 elements t_1=1, t_2=1, t_3=0, t_4=0, t_5=1, t_6=0, t_7=1, t_8=1, t_9=0, t_{10}=0, t_{11}=0, t_{12}=1.

Therefore the sum $x = \sum_{1\leq i\leq 12} w_i.t_i = w_1\oplus w_2 \oplus w_5\oplus w_7\oplus w_8\oplus w_{12}$=000101.

Now, suppose we need to hide a 5-bit string D_0 = 10110 in S:

- We extend D_0 to D = 101101;
- Since $x \neq D$ and $y = D \oplus x$ =$101000_{(2)}$=40, we can not find w_i such that w_i =40 in U, We find w_{10}= 6 and w_{12}=32 in U such that 8⊕32 = 40.
  Then we flip p_{10} to 0 and p_{12} to 0 in S, we get S (new) with p_{10}=0 and p_{12} =0.

In the extracting phase, we calculate T= [$S \oplus K$] which has 12 elements t_1=1, t_2=1, t_3=0, t_4=0,t_5=1,t_6=0, t_7=1, t_8=1, t_9=0, t_{10}=1, t_{11}=0, t_{12}=0. Then we compute $x = \sum_{1 \leq i \leq 12} w_i.t_i = w_1 \oplus w_2 \oplus w_5 \oplus w_7 \oplus w_8 \oplus w_{10}$ = 101101. By x is odd, this informs that S is a success block, hence we can extract D_0= x div 2 = 10110 as claimed.

3.4 Hiding Data in a Sequence of Blocks of an Binary Image G

Since the size 4 × 3 is small, we can use a set K of 12 × 10 =120 binary key values and split K into 10 binary matrices as K_0, K_1,.., K_9, each of them has the same size 4×3. Suppose we need to hide a sequence of 5-bit strings D_1,..,D_q in a sequence of blocks S_1,..,S_N of a binary image G. We present two hiding and extracting phases as follows.

3.4.1 Hiding Phase

```
Step1)  Set $t$ = 0; $j$=1;$i$=1;
Step 2) While ($i$ <= $q$) and ($j$<=$N$) do
            2.1) Try to hide $D_i$ in $S_j$ by algorithm B1 using the matrix key $K_t$ ;
            2.2) If $S_j$ is a success block, set
                    $i$=$i$+1; $j$=$j$+1; $t$ = ($t$+1) mod 10;
            2.3.) If $S_j$ is failure, set $j$=$j$+1;
        End while;
Step 3) If ($i$>$q$)  then Return(G) as stego image
        Else Return($i$-1); {inform in $G$ we can hide successfully only $i$ items $D_1$,..,$D_i$,
not all the sequence $D_1$,..,$D_q$}
```

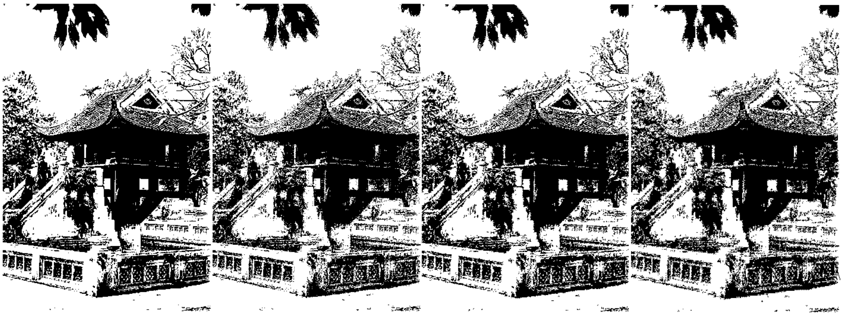
3.4.2 Extracting Phase

```
Step1)  Set $t$ = 0; $j$=1;$i$=1;
Step 2) While ($i$ <= $q$) and ($j$<=$N$) do
            2.1) Try to extract $D_i$ in $S_j$ by algorithm B2 using the matrix key $K_t$ ;
            2.2) If $S_j$ is a success block, $D_i$ is saved and set
                    $i$=$i$+1; $j$=$j$+1; $t$ = ($t$+1) mod 10;
            2.3.) If $S_j$ is failure, set $j$=$j$+1;
        End while;
Step 3) If ($i$>$q$) then Return(sequence $D_1$,..,$D_q$) as embedded data
        Else  inform that $G$ can not used successfully to extract all the sequence
$D_1$,..,$D_q$ and Exit.
```

4 Experimental Results

We build a program which gives us many 2-weak bases in $\mathbf{Z}_2^6$ for applications to binary and palette images. For example, another 2-weak base is V= {11, 51, 55, 46, 41, 29,1,2,4,8,16,32}. The unique element 58 in M can not be presented by V.

We can deverlope the way to apply Parity Approach (PA) due to our previous work [7] or works in [4,5] to build a new schemes applied for palette images with a modification from the scheme in section 3. At first, given a palette image G with palette P, by the parity values Val(p) of all pixels p of G, where Val: $P \rightarrow \mathbf{Z}_2$ is the function defined as in our works [7] or [4,5], we can build a sequence S_1,S_2,..,S_N of binary blocks from G. For each block, we can hide a sequence of 6-bit strings in, except that in case we need flip a pixel p* in each S_i, we flip the color of the compative pixel p in G to the closed colors p'=Next(p) in the palette P by a Next function. The Next and Val are mentioned as in (2.1) and (2.2) conditions in the section 2 above. For grayscale images, Next(p) can be defined simply as p+1 or p-1. The following binary images in Fig.1 are stego-images of size 350 x 514, each image is spitted in to blocks of size 4x3, these are obtained by CPT [3], MCPT [10] schemes and our proposed scheme. The results show the highest capacity of our scheme with reasonable quality (by PSNR). This scheme can be modified for grayscale and palette images to obtain hiding schemes with high quality. By our results, for stego-gray images of 256 levels, size 512 x 512, the number of hidden bytes is 16383 (all blocks are success), all PSNR's obtained are larger than 56 dB.



One-pillar pagoda of Hanoi, original binary image, size 350 x 514

Hide 2222 bytes by CPT scheme PSNR 63.1 dB

Hide 1376 bytes by MCPT scheme PSNR 63.67 dB

Our scheme- hide 3251 bytes PSNR 61.75 dB

Fig. 1. One-Pillar Pagoda of Hanoi- hiding data by our and CPT, MCPT schemes

5 Conclusion

The paper presents a new hiding scheme to hide a large number of bits in binary images by small blocks while the quality can be controlled with a large set of binary keys for security claims, by applying properties of module over $\mathbf{Z}_2$. For real applications, this scheme can be modified easily to grayscale, palette images to obtain hiding schemes with high quality. In our future works, images in other format or audio, video can be considered to get good hiding schemes based on modules.

References

1. Bierbrauer, J., Fridrich, J.: Constructing Good Covering Codes for Applications in Steganography. In: Shi, Y.Q. (ed.) Transactions on DHMS III. LNCS, vol. 4920, pp. 1–22. Springer, Heidelberg (2008)
2. Calderbank, A.R., Sloane, N.J.A.: Inequalities for covering codes. IEEE Transactions on Information Theory IT-34, 1276–1280 (1988)
3. Chen, Y., Pan, H., Tseng, Y.: A secret of data hiding scheme for two-color images. In: IEEE Symposium on Computers and Communications (2000)
4. Romana Machado. EZStego[EB/OL], `http://www.stego.com`
5. Fridrich, J., Du, R.: Secure Steganographic Methods for Palette Images. In: Pfitzmann, A. (ed.) IH 1999. LNCS, vol. 1768, pp. 47–60. Springer, Heidelberg (2000)
6. Phan, T.H., Nguyen, H.T.: On the Maximality of Secret Data Ratio in CPTE Schemes. In: Nguyen, N.T., Kim, C.-G., Janiak, A. (eds.) ACIIDS 2011, Part I. LNCS (LNAI), vol. 6591, pp. 88–99. Springer, Heidelberg (2011)
7. Huy, P.T., Thanh, N.H., Thang, T.M., Dat, N.T.: On Fastest Optimal Parity Assignments in Palette Images. In: Pan, J.-S., Chen, S.-M., Nguyen, N.T. (eds.) ACIIDS 2012, Part II. LNCS (LNAI), vol. 7197, pp. 234–244. Springer, Heidelberg (2012)
8. Zhang, X., Wang, S.: Vulnerability of pixel-value differencing steganography to histogram analysis and modification for enhanced security. Pattern Recognition Letters 25, 331–339 (2004)
9. Zhang, X., Wang, S.: Analysis of Parity Assignment Steganography in Palette Images. In: Khosla, R., Howlett, R.J., Jain, L.C. (eds.) KES 2005, Part III. LNCS (LNAI), vol. 3683, pp. 1025–1031. Springer, Heidelberg (2005)
10. Tseng, Y.-C., Pan, H.-K.: Secure and InvisibleData Hiding in 2-Color Images. In: Proceedings of INFOCOM 2001, pp. 887–896 (2001)

Runtime Verification of Multi-agent Systems Interaction Quality

Najwa Abu Bakar and Ali Selamat

Faculty of Computer Science and Information System
Universiti Teknologi Malaysia
81300, Skudai,
Johor Darul Takzim
najwa.abakar@gmail.com, aselamat@utm.my

Abstract. Since multi-agent systems are inherently complex, there are possibilities that errors related to multi-agent systems interaction could occur. Currently, many verification approaches have been proposed by focusing on specific properties, using a particular technique and during certain development phase. However, each technique has its limitations. As interaction between agents and multi-agent systems environments evolve during runtime, not all multi-agent systems interaction requirements can be specified and verified during design and development. Thus, some new interaction properties such as agent availability and trustability need to be verified during runtime. In this research, a solution is proposed in which newly defined agents interaction quality requirements are specified, developed into metrics and verified within multi-agent systems runtime verification framework. It is aimed to improve the effectiveness of the verification of agent interactions during runtime. Finally, an experiment is set up to capture message passing between agents and to gather runtime system profiles to evaluate the proposed solution.

Keywords: Multi-agent systems, runtime verification, interaction quality, message passing.

1 Introduction

Existing MAS verification approaches [36] can be divided into three stages that are verification during 1) design, 2) development and 3) runtime. Although the approaches, that have been properly designed, tested using case studies and evaluated, have marginally improved correctness of MAS, each level of approaches has its own limitations and a lot of improvements are still needed. First, for verification during design, formal verification is performed automatically using model checking [1][5][8][29], automated theorem proving [32], or simulation [21]. As MAS is inherently complex [11], performing model checking only manages to verify correctness of certain properties [7][16][30]. Besides, as the model checking rely on the input which are properties specification and design model [9][15][17][20], the quality of these two inputs contribute to the accuracy of the checking. Coverage or

A. Selamat et al. (Eds.): ACIIDS 2013, Part I, LNAI 7802, pp. 435–444, 2013.

degree of thoroughness of the specified properties and accuracy of the modeling still need to be measured in order to assess and increase the accuracy of the model checking output. Second, similarly, during development [22], MAS debugging and testing [28] also suffer from the incomprehensiveness of the data captured during monitoring and the thoroughness of the test suites. Finally, verification during runtime [2][12][24][27] that focuses only towards agents, integration, and message passing without considering supporting contextual information also faces similar issues that are incompleteness of the data to be analyzed [4].

The above discussed problems highlight the current MAS correctness verification issues that need to be improved by tackling the comprehensiveness and pre-processing or data preparation issues of the verification. For the MAS that has been developed and executed, the only way to assess its correctness is by performing analysis towards MAS interaction and communication activities. The richness of the data captured during those activities has opened the opportunity for low-level analysis of the data, i.e. interaction messages that can be manipulated and prepared to fulfill the gap between low-level MAS infrastructure and the verification and analysis at higher-level. Observing this issue from software quality point of view, it is believed that the incomprehensive verification issue faced by existing verification approaches can be improved by considering the MAS interaction resources data and data captured during interaction activities. Implementing the framework during runtime complements the existing verification performed during design [3] using model checking or theorem proving [1][5][8][29][32] in which new requirements can be defined and verified after the systems have been executed [23][33].

In this paper, we propose a framework called Multi-agent Runtime Verification (MARV) framework. In this framework, we define new requirements of MAS interaction during runtime and develop new metrics to improve the effectiveness of MAS interaction verification. This framework will be implemented in order to evaluate the effectiveness of the metrics in improving MAS interaction error detection rate. Two interaction qualities to be verified are defined in this paper that are the agents *availability* and *trustability*.

The rest of the paper is organized as follows. Section 2 provides the research background and analysis of the existing works. Section 3 describes the proposed solution that is the requirements definition process within MARV framework and in Section 4, we describe the implementation and evaluation consideration. Finally, in Section 5, we present the discussion and future work.

2 Background and Related Works

2.1 MAS Interaction Verification

There are three levels of approaches to improve correctness of MAS interactions that are design level, development level and runtime level. During design, verification of MAS designs are performed against specified interaction properties and during development, debugging and testing are performed to find bugs. During runtime, there are two main components implemented that are monitoring and runtime verification [12][27]. During design, many studies have been performed in the area of

MAS verification. There are many approaches proposed. One of them is by inventing tools e.g. MCMAS (model checker) [25][29] and SOCS-SI (automated theorem proving) [32] customized to perform formal verification on MAS during design time by considering common agent abstractions. Another approach is by checking on selected critical smaller models extracted from the complex MAS and use general-purpose distributed model checking tools such as SPIN, UPPAAL, etc to verify against general properties such as temporal logic [3]. Both mentioned approaches are performed during design time. The framework proposed in this paper aims to complement these approaches. New interaction requirements are defined after the systems are executed based on MAS agents and environment contextual information.

2.2 Existing MAS Interaction Requirements

From the literature, MAS requirements can be divided into two categories that are MAS functional requirements and non-functional requirements. Functional requirements for a system are the definition of the behavior or functions of the systems, what the system "shall do" (behavior) while non-functional requirements define how the system "shall be" (constraints). The functional requirements are explained in the system design while the non-functional requirements are detailed out in the system architecture [13][19]. *Functional requirements* for agents are the characteristics of agents that are autonomous, reactive, proactive, and coordinating based on agents roles, types, goals, and tasks assigned to them. For example, mobile agents are able to migrate from one platform to another to achieve goals and complete tasks. *Non-functional requirements* are based on the constraints of the agents such as safety, integrity and quality of agents to protect and to ensure that the agents are reliable and can be trusted [31][34][35]. Examples of non-functional requirements of MAS are:

- Compatibility; the ability to follow the standard specified by FIPA [18] and application ontology.
- Interoperability; the ability to work with other agents and applications
- Safety; the ability of the agents to protect themselves from threats.
- Credibility; trust and reputation of agents.
- Integrity; to ensure that messages between agents, between agents and agent platform, and between agents with environment are not altered.
- Availability; to ensure services are not suffered from denial of service (DoS)
- Confidentiality; to ensure that only certain agents with certain roles, trusted levels, or reputation are allowed to access information.

Investigation towards existing works in MAS verification shows that these MAS requirements are specified as MAS properties that can be classified into interaction, behavior, and knowledge properties for verification. In this research, we are focusing on MAS interaction issues.

3 Requirements Definition Process in MARV

3.1 Definition Process

In the requirements definition phase in MARV, a few steps are performed: 1) Defining Agents Interaction Quality Criteria, 2) Defining the Agents Interaction Data, 3) Defining Agents Interaction Quality Requirements and Parameters, and 4) Defining the Agents Interaction Quality Rules and Metrics. These steps are explained in the following subsections.

3.2 Defining Agents Interaction Quality Criteria

The first task in designing MARV is to determine what factors constitute a successful agent conversation or interaction. The quality of agents interaction here means the success of the conversation according to the specifications and application's contextual condition. In other words, the selected criteria are preferably in measurable form that can be used to quantify a message as having the characteristics to be successful in agent conversation. A good start in this direction is to refer to other works in data quality, such as Total Quality Management (TQM) [10] and Information Quality (IQ) [26]. These studies suggested that data quality can be defined as combination of scores from multiple data quality parameters. Some of the commonly used data quality parameters are accuracy, completeness, relevancy, timeliness and credibility. These are general parameters to describe the criteria for information or product. In this case, these general criteria have to be adapted and shaped towards the area of communication of multi-agent systems. Whatever criteria that are eventually selected, they need to be accurate and comprehensively selected since they represent the quality of the conversation messages in general. Furthermore, the metrics (or scores) generated for these criteria will determine the accuracy of the verification result, which has the direct impact in classifying messages into valid or invalid messages during analysis stage. Figure 1 shows the verification process cycle adapted from the TQM.

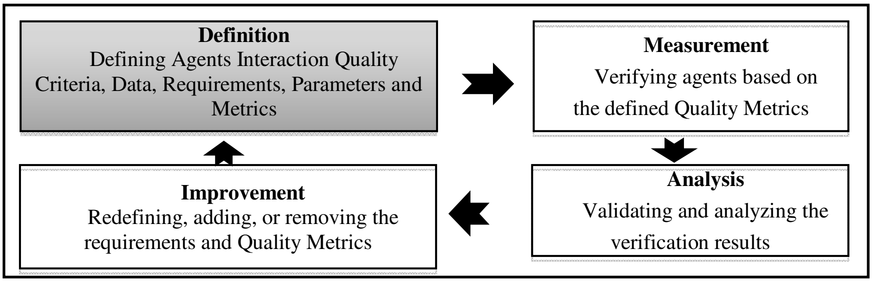

Definition
Defining Agents Interaction Quality Criteria, Data, Requirements, Parameters and Metrics
Measurement
Verifying agents based on the defined Quality Metrics
Analysis
Validating and analyzing the verification results
Improvement
Redefining, adding, or removing the requirements and Quality Metrics

Fig. 1. The verification process cycle of the MARV components adapted from the TQM. The definition process is highlighted in grey box.

3.3 Defining the Agents Interaction Data – The Messages

The raw data to be verified in MARV is the set of agents interaction "messages". A message is the basic element that forms conversation during agents communication. Hence, it is logical to start the definition stage by looking at the characteristics of a typical message. Table 1 below shows a sample message labeled with the attributes on the left column that are AgentId, ConversationId, Sender, Receiver, Performative, Content, Language, Ontology, Protocol, and Reply-by.

Table 1. Sample of agents interaction message

Message Attributes	Example
AgentId	1
ConversationId	1
Sender	ipa@mydomain.com
Receiver	aa@yourdomain.com
Performative	request
Content	""((action (AgentId :name coordinator@yourdomain.com)(contact-student :offer-study-level Master Degree :course Computer Science :registration 01/09/2012)))""
Language	FIPA-SL
Ontology	university-online-application
Protocol	fipa-request
Reply-by	null

3.4 Defining Agents Interaction Quality Requirements

In order to verify the agent interaction messages, the general data quality requirements have been identified to show the validity of a message as depicted below. There are many other requirements but here, the most important, relevant, and related to agents, platforms, hosts, environments or networks, and devices/hardware (contextual information) are identified. The list is not exhaustive; rather it should be considered as the minimal interaction correctness requirements for interaction verification. The list can be extended further as the systems profile, knowledgebase and configuration evolve during runtime. The requirements, R1, R2 and R3 are listed below:

R1: The receiver agent is available during message transmission. The assumption here is that if the receiver is available, there is high chance that the message will be received successfully compared to unavailable or busy receiver agent.

R2: The receiver agent is authorized to receive the content of INFORM message. Here, it is assumed that critical information can only be received by agents that have the authority. Only certain agents with specific roles and located in trusted location (platform, host, and network) can receive certain information.

R3: The sender agent is trusted to send an INFORM or a REQUEST message. Only certain agents with specific roles and located in trusted location (platform, host, or environment) can send certain information and request.

The abstraction of *agent interaction quality requirements* stated above can be translated into a more concrete form of representation known as *agent interaction quality metrics.* Rules and indicators (in the form of scores) are associated for each parameter based on the related conditions, and the scores. In the following subsection, agents interaction quality metrics for each interaction quality requirements are defined.

3.5 Defining the Agents Interaction Quality Rules and Metrics

A data quality metric is a measure of some property of a piece of information or its specifications. In order to measure the agents interaction quality, based on the abstract agents interaction requirements defined above, in this phase, the agents interaction quality rules are constructed. The message correctness requirements identified during definition process are transformed into rules and measurable formulations or indicators (in the form of scores) that reflect the specifications and real condition (supported by contextual information) of the messages. Table 2 presents the proposed rules for each interaction quality requirement. The rules are constructed based on the interaction quality requirements for interaction verification identified in Section 3.4 above.

The rule for each agents interaction metric is a simple if-then-else statement typical in many programming algorithm. Basically, for each metric, the rule checks via its if-then-else statement, the message's attributes value against the monitored information in agent profiles and MAS knowledgebase resources, and then assign score accordingly as shown in Table 2 below.

Table 2. The constructed agents interaction rules and metrics

Require-ments	Metrics	Message Attributes	Rules
Layer 1: Agent Verification Level			
R1	Availability-receiver (AR)	Receiver	For each message, if the receiver agent is available to receive a message, score=1 else score=0.
R2	Trustability-receiver (TR)	Performative Receiver	For INFORM message, if the receiver agent is authorized, score=1 else score=0.
R3	Trustability-sender (TS)	Performative Sender	For INFORM message, if the sender agent is trusted, score=1 else score=0.

As shown in Table 2, the verification of sender and receiver agents includes the checking of receiver agent's availability and trustability of the sender and receiver. Agents as the main players in MAS interactions are the main factors of the success of the interaction and also the main sources for the communication errors to occur. Thus, agents are required to be available and trustable during interaction.

4 Implementation and Evaluation Consideration

The definition processes are implemented within MARV framework. The architecture is shown in Figure 2 below. MARV consists of four main verification process components: Definition, Measurement, Analysis, and Improvement adapted from TQM (refer Section 3.2). In addition, MARV also includes the additional components that are the *Agent Communication Language* (*ACL) messages capturing*, *Agents and MAS infrastructure profiling* and *MAS Knowledgebase gathering*.

During ACL messages capturing, the collected message's attributes are transformed into database fields. The database fields include: SessionId, MessageId, Timestamp, Valid, Size, Late, ConversationId, Sender, Receiver, Performative, Content, Language, Ontology, Protocol, and Reply-by. Next, to gather agents and MAS infrastructure profiles, the list of agents and the attributes related to each agent and its infrastructure are stored in database fields. The database fields include: SessionId, AgentId, and AgentName. In this research, the tables are expanded by adding several more database fields to provide extra information related to the infrastructure, environment, and location of the agents: Container, Platform, Host, and Network. Finally, for MAS knowledgebase gathering, there are two main sources of the knowledgebase that are the ontologies and interaction protocols. Both the MAS's specified ontology and used interaction protocol are identified and the information are extracted from the coding and FIPA standard website to be stored in database fields.

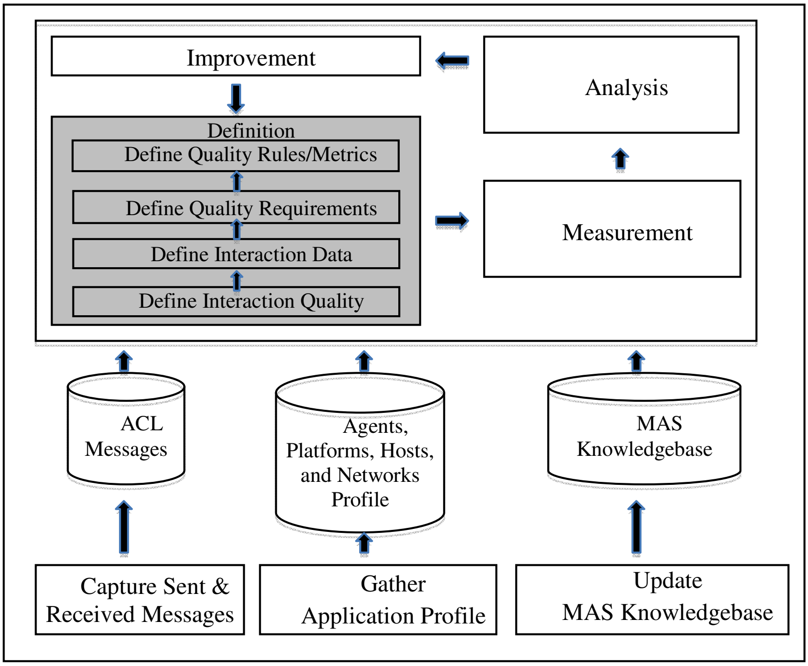

Improvement
Analysis
Definition
Define Quality Rules/Metrics
Define Quality Requirements
Define Interaction Data
Define Interaction Quality
Measurement
ACL Messages
Agents, Platforms, Hosts, and Networks Profile
MAS Knowledgebase
Capture Sent & Received Messages
Gather Application Profile
Update MAS Knowledgebase

Fig. 2. The definition processes within MARV framework architecture (highlighted in grey boxes)

To test the framework, experiments are performed. Given a set of exchanged messages, considering the MAS knowledgebase and infrastructure profiling, benchmark data are prepared by manually classifying the messages into correct and incorrect messages. The incorrect messages identified during this manual classification stage is called *known incorrect messages*. These known incorrect messages are the messages that contain errors supposed to be detected using MARV during runtime. From the known incorrect messages, *known unique errors* are extracted. These errors are supposed to be detected by the defined metrics and rules (properties). Next, using the MARV, the messages are verified during runtime and classified automatically into correct and incorrect messages. The identified suitable evaluation metrics to measure the effectiveness of the MARV tool in performing MAS interaction verification are *precision* and *recall*, *properties coverage* and *time* taken to perform the verification.

5 Discussion and Future Work

We have presented the MARV, a runtime multi-agent verification framework to verify MAS interaction. This effort to detect agents message passing errors during runtime can increase the effectiveness of MAS verification. This is because the correctness of the MAS not only considers design issues but also other runtime factors such as agents profile, knowledgebase and configuration that can evolve during runtime. The experiment is set up using JADE [6][14] environment as one of the case studies. It is to implement the MARV framework that includes the capturing of message passing between agents and the gathering of runtime system profiles to evaluate the proposed solution. In the future, the presented rules and metrics presented in this study will be properly developed into mathematical formula that will be used to design a verification algorithm.

Acknowledgments. The authors thank Ministry of Higher Education Malaysia (MOHE) under Fundamental Research Grant Scheme (FRGS) Vot 4F031 and Universiti Teknologi Malaysia under Research University Funding Scheme (Q.J130000.7110.02H47) for supporting the related research.

References

1. Latif, N.A., Hassan, M.F., Hasan, M.H.: Formal Verification for Interaction Protocol in Agent-Based E-Learning System Using Model Checking Toolkit - MCMAS. In: Zain, J.M., Wan Mohd, W.M.b., El-Qawasmeh, E. (eds.) ICSECS 2011, Part II. CCIS, vol. 180, pp. 412–426. Springer, Heidelberg (2011)
2. Abdul Bujang, S., Selamat, A.: Verification of Mobile SMS Application with Model Checking Agent. In: Proceedings of the 2009 International Conference on Information and Multimedia Technology, ICIMT 2009, pp. 361–365. IEEE Computer Society, Washington, DC (2009)

3. Abu Bakar, N., Selamat, A.: Analyzing model checking approach for multi agent system verification. In: 2011 5th Malaysian Conference on Software Engineering, MySEC, pp. 95–100 (2011)
4. Abu Bakar, N., Selamat, A.: Towards Implementing Dynamic Multi-agent V&V Framework. In: The Third Software Engineering Postgraduates Workshop, SEPoW 2011, JB, Malaysia (2011)
5. Alechina, N., Logan, B., Nguyen, H.N., Rakib, A.: Automated Verification of Resource Requirements in Multi-Agent Systems Using Abstraction. In: van der Meyden, R., Smaus, J.-G. (eds.) MoChArt 2010. LNCS, vol. 6572, pp. 69–84. Springer, Heidelberg (2011)
6. Bellifemine, F.L., Caire, G., Greenwood, D.: Developing multi-agent systems with JADE. John Wiley & Sons, Ltd., West Sussex (2007)
7. Ben Ayed, L.J., Siala, F.: Event-B based Verification of Interaction Properties In Multi-Agent Systems. Journal of Software 4(4), 357–364 (2009)
8. Benerecetti, M., Cimatti, A.: Validation of Multiagent Systems by Symbolic Model Checking. In: Giunchiglia, F., Odell, J.J., Weiss, G. (eds.) AOSE 2002. LNCS, vol. 2585, pp. 32–46. Springer, Heidelberg (2003)
9. Berard, B., Bidoit, M., Finkel, A., Laroussinie, F., Petit, A., Petrucci, L., et al.: Systems and Software Verification: Model-Checking Techniques and Tools (1999)
10. Besterfield, D., Besterfield-Michna, C., Besterfield, G., Besterfield-Sacre, M., Urdhwareshe, H., Urdhwareshe, R.: Total Quality Management. Pearson Education (2011)
11. Bordini, R., Dastani, M., Dix, J., Seghrouchni, A.: Multi-Agent Programming: Languages, Tools and Applications. Springer, NY (2009)
12. Botía, J.A., Gómez-Sanz, J.J., Pavón, J.: Intelligent Data Analysis for the Verification of Multi-Agent Systems Interactions. In: Corchado, E., Yin, H., Botti, V., Fyfe, C. (eds.) IDEAL 2006. LNCS, vol. 4224, pp. 1207–1214. Springer, Heidelberg (2006)
13. Burnstein, I.: Practical Software Testing: A Process-Oriented Approach. Springer, NY (2003)
14. Caire, G., Pieri, F.: JADE. Java Agent Development Framework (2011), `http://jade.tilab.com/doc/tutorials/LEAPUserGuide.pdf` (retrieved May 18, 2012)
15. Clarke, E.M., Grumberg, O., Peled, D.A.: Model Checking. The MIT Press, Cambridge (1999)
16. Dekhtyar, M.I., Dikovsky, A.J., Valiev, M.K.: Temporal Verification of Probabilistic Multi-Agent Systems. In: Avron, A., Dershowitz, N., Rabinovich, A. (eds.) Pillars of Computer Science. LNCS, vol. 4800, pp. 256–265. Springer, Heidelberg (2008)
17. Dennis, L., Fisher, M., Matthew, W.P., Bordini, R.H.: Model Checking Agent Programming Languages. Automated Software Engineering 19(1), 5–63 (2012)
18. FIPA.: The Foundation for Intelligent Physical Agents (2012), `http://www.fipa.org/` (retrieved May 18, 2012)
19. Fulcher, J.: Advances in Applied Artificial Intelligence. Idea Group Publishing, London (2006)
20. Gammie, P., van der Meyden, R.: MCK: Model Checking the Logic of Knowledge. In: Alur, R., Peled, D.A. (eds.) CAV 2004. LNCS, vol. 3114, pp. 479–483. Springer, Heidelberg (2004)
21. Giese, H., Klein, F.: Systematic verification of multi-agent systems based on rigorous executable specifications. International Journal of Agent-Oriented Software Engineering 1(1), 28–62 (2007)
22. Gómez-Sanz, J.J., Botía, J., Serrano, E., Pavón, J.: Testing and Debugging of MAS Interactions with INGENIAS. In: Luck, M., Gomez-Sanz, J.J. (eds.) AOSE 2008. LNCS, vol. 5386, pp. 199–212. Springer, Heidelberg (2009)

23. Hallé, S., Villemaire, R.: Runtime Verification for the Web: A Tutorial Introduction to Interface Contracts in Web Applications. In: Barringer, H., Falcone, Y., Finkbeiner, B., Havelund, K., Lee, I., Pace, G., Roşu, G., Sokolsky, O., Tillmann, N. (eds.) RV 2010. LNCS, vol. 6418, pp. 106–121. Springer, Heidelberg (2010)
24. Selamat, A., Lockman, M.T.: Multi-agent Verification of RFID System. In: Nguyen, N.T., Katarzyniak, R.P., Janiak, A. (eds.) New Challenges in Computational Collective Intelligence. SCI, vol. 244, pp. 255–268. Springer, Heidelberg (2009)
25. Lomuscio, A., Qu, H., Raimondi, F.: MCMAS: A Model Checker for the Verification of Multi-Agent Systems. In: Bouajjani, A., Maler, O. (eds.) CAV 2009. LNCS, vol. 5643, pp. 682–688. Springer, Heidelberg (2009)
26. Naumann, F.: Quality-Driven Query Answering. LNCS, vol. 2261. Springer, Heidelberg (2002)
27. Osman, N.: Runtime Verification of Deontic and Trust Models in Multiagent Interactions. Phd Thesis (2008)
28. Poutakidis, D.: Debugging Multi-Agent Systems With Design Document. Phd Thesis (2008)
29. Raimondi, F.: Model checking multi-agent system. Phd Thesis (2006)
30. Sabri, K.E., Khedri, R., Jaskolka, J.: Verification of Information Flow in Agent-Based Systems. In: Babin, G., Kropf, P., Weiss, M. (eds.) MCETECH 2009. LNBIP, vol. 26, pp. 252–266. Springer, Heidelberg (2009)
31. Silva, C., Pinto, R., Castro, J., Tedesco, P.: Requirements for Multi-Agent Systems. In: Workshop em Engenharia de Requisitos, WER, Piracicaba-SP, Brasil, pp. 198–212 (2003)
32. Singh, M.P., Chopra, A.K.: Correctness Properties for Multiagent Systems. In: Baldoni, M., Bentahar, J., van Riemsdijk, M.B., Lloyd, J. (eds.) DALT 2009. LNCS, vol. 5948, pp. 192–207. Springer, Heidelberg (2010)
33. Stoller, S.D., Bartocci, E., Seyster, J., Grosu, R., Havelund, K., Smolka, S.A., Zadok, E.: Runtime Verification with State Estimation. In: Khurshid, S., Sen, K. (eds.) RV 2011. LNCS, vol. 7186, pp. 193–207. Springer, Heidelberg (2012)
34. Sycara, K.P.: Multiagent Systems. AI Magazine 19(2), 79–92 (1998)
35. Wooldridge, M.: An Introduction to MultiAgent Systems, 2nd edn. John Wiley & Sons, United Kingdom (2009)
36. YAHODA.: Verification Tools Database (2011), `http://anna.fi.muni.cz/yahoda/` (retrieved May 18, 2012)

Bounds on Lengths of Real Valued Vectors Similar with Regard to the Tanimoto Similarity

Marzena Kryszkiewicz

Institute of Computer Science, Warsaw University of Technology
Nowowiejska 15/19, 00-665 Warsaw, Poland
mkr@ii.pw.edu.pl

Abstract. The Tanimoto similarity measure finds numerous applications in chemistry, bio-informatics, information retrieval and text mining. A typical task in these applications is finding most similar vectors. The task is very time consuming in the case of very large data sets. Thus methods that allow for efficient restriction of the number of vectors that have a chance to be sufficiently similar to a given vector are of high importance. To this end, recently, we have derived bounds on lengths of vectors similar with respect to the Tanimoto similarity. In this paper, we recall those results and derive new bounds on lengths of real valued vectors that have a chance to be Tanimoto similar to a given vector in a required degree. Finally, we compare the previous and current results and illustrate their usefulness.

Keywords: the Tanimoto similarity, chemical substructure discovery, text mining, information filtering.

1 Introduction

The Tanimoto similarity measure finds numerous applications in chemistry, biology, bio-informatics, information retrieval and text mining [1, 8]. A typical activity in these applications is finding most similar vectors representing respective objects such as chemical substructures or documents. For instance, documents are often represented as term frequency vectors or its variants such as tf_idf vectors [7]. In the case of binary vectors, the Tanimoto similarity between two vectors equals the ratio of the number of attributes with “1s” shared by both vectors to the number of attributes with “1s” that occur in either vector. In this particular case of vectors, the Tanimoto similarity is equivalent to the Jaccard similarity, which is defined for pairs of sets as the ratio of the cardinality of the intersection of the two sets to the cardinality of their union. However, the Tanimoto similarity be calculated for any non-zero vectors and thus is regarded as an extension of the Jaccard similarity.

The determination of Tanimoto similar vectors is very time consuming in the case of very large data sets. In the case of sparse high dimensional datasets (with thousands or tens of thousands of dimensions) such as text datasets (representing documents or contents of Web pages), inverted indices are used to cope with this problem [1, 9]. The triangle inequality can be also used to mitigate the problem [2, 3, 5, 10]. However, any new

A. Selamat et al. (Eds.): ACIIDS 2013, Part I, LNAI 7802, pp. 445–454, 2013.

methods that would enable even more powerful restriction of the number of vectors that must be evaluated as potentially sufficiently similar to a given vector are still of high importance. In the case of Jaccard measure, one may apply to this end bounds on cardinalities of similar sets [6]. Thanks to the equivalence between the Jaccard similarity and the Tanimoto similarity between binary vectors, this result can be easily adapted for looking for Tanimoto similar binary vectors. Recently, we have proposed in [4] bounds on lengths of Tanimoto similar vectors in the case of vectors with specific two-valued or three-valued dimensions as well as for real valued vectors. The formulae expressing the bounds in the latter case were obtained in an arduous way, and were quite complicated. To the best of our knowledge, no other bounds on lengths of Tanimoto similar real valued vectors were offered so far. In this paper, we propose new bounds on Tanimoto similar real valued vectors and compare them with the results obtained in [4].

Our paper has the following layout. Section 2 provides basic notions and properties related to the Tanimoto similarity. In Section 3, we describe related work on bounds on lengths of Tanimoto similar vectors. In Section 4, we propose and prove new bounds on lengths of Tanimoto similar real valued vectors. In Section 5, we analyze the recalled and the proposed bounds and conclude our findings. Experimental results are provided in Section 6. Section 7 summarizes our work.

2 Basic Notions and Properties

In the paper, we will consider n-dimensional vectors. A vector u will be also denoted as $[u_1, \ldots, u_n]$, where u_i is the value of the i-th dimension of u, $i = 1..n$. A vector will be called a *non-zero vector* if at least one of its dimensions has non-zero value.

A popular form of vectors, especially in text mining and chemistry, are binary vectors, each domain of which may take either value 0 or 1, where 1 denotes presence of an attribute, while 0 its absence. If a two-valued domain of a vector contains a positive value different from 1, then this might reflect the importance of the occurrence of the attribute. However, vectors with three-valued dimensions also may find applications for example in looking for documents that cite similar papers in a similar way: if a paper is cited as valuable, it could be graded with a positive value; if it is cited as invaluable, it could be graded with a negative value; if it is not cited, it could be graded with 0 [4].

Similarity between vectors is often measured by means of the the *Tanimoto similarity*. The *Tanimoto similarity* between vectors u and v will denoted by $T(u, v)$ and will be defined as follows,

$$T(u, v) = \frac{u \cdot v}{u \cdot u + v \cdot v - u \cdot v}$$

where $u \cdot v = \sum_{i=1..n} u_i v_i$.

Property 1. $T(u, v) = \dfrac{u \cdot v}{| u |^2 + | v |^2 - u \cdot v}$

where $| u |$ is *the length of vector* u and equals $\sqrt{u \cdot u}$.

Property 2 [8]. Let u and v be non-zero vectors. Then, $T(u, v) \in \left[-\frac{1}{3}, 1\right]$.

In this paper, given length of a vector u, we will focus on deriving bounds on lengths of real valued vectors v such that $T(u, v) \geq \varepsilon$.

3 Related Work

As we have already mentioned, the Jaccard measure coincides with the Tanimoto measure for binary vectors. We will recall now the definition of the Jaccard similarity to show this correspondence. The *Jaccard similarity* between sets U and V will be denoted by $J(U, V)$ and will be defined as follows,

$$J(U, V) = \frac{|U \cap V|}{|U \cup V|} = \frac{|U \cap V|}{|U| + |V| - |U \cap V|}.$$

Beneath we recall the bounds on cardinalities of sets similar with respect to the Jaccard measure based on [6].

Property 3. Let U and V be non-empty sets, $J(U, V) \geq \varepsilon$ and $\varepsilon \in (0, 1]$. Then:

$$|V| \in \left[\varepsilon |U|, \frac{|U|}{\varepsilon}\right].$$

Clearly, any set $U \subseteq \Omega$, where $|\Omega| = n$, can be represented uniquely as the binary n-dimensional vector u such that its i-th dimension equals 1 if i-th element of Ω belongs to U and 0, otherwise. Obviously, if u and v are such binary n-dimensional vectors representing sets U and V, then $u \cdot v = |U \cap V|$, $|u|^2 = |U|$, $|v|^2 = |V|$, so, $T(u\ v) = J(U, V)$. Hence, the bounds on lengths of binary vectors similar with respect to the Tanimoto similarity are obtainable immediately from the bounds on cardinalities of sets similar with respect to the Jaccard measure.

Corollary 1. Let u and v be non-zero binary vectors such that $T(u, v) \geq \varepsilon$ and $\varepsilon \in (0, 1]$. Then:

$$|v|^2 \in \left[\varepsilon |u|^2, \frac{|u|^2}{\varepsilon}\right],$$

or equivalently,

$$|v| \in \left[\sqrt{\varepsilon}\, |u|, \frac{|u|}{\sqrt{\varepsilon}}\right].$$

In our recent work [4], we have generalized the above result for specific vectors with two-valued and/or three-valued domains as follows:

Theorem 1 [4]. Let *u* and *v* be non-zero vectors each domain V_i, i = 1..*n*, of which equals either {0, a_i}, where a_i is a real value, or {0, a_i, b_i}, where $a_i\, b_i < 0$ and a_i, b_i are real values. If $T(u, v) \geq \varepsilon$ and $\varepsilon \in (0, 1]$, then:

$$|v| \in \left[\sqrt{\varepsilon}\,|u|, \frac{|u|}{\sqrt{\varepsilon}}\right].$$

In addition, we have proposed and proved in [4] the bounds on lengths of any non-zero vectors similar with respect to the Tanimoto similarity for $\varepsilon \in \left(\frac{1}{3}, 1\right)$ as follows.

Theorem 2 [4]. Let *u* and *v* be non-zero vectors such that $T(u, v) \geq \varepsilon$, $\varepsilon \in \left(\frac{1}{3}, 1\right)$,

$$\alpha = \frac{1}{2}\left(\left(1+\frac{1}{\varepsilon}\right) + \sqrt{\left(1+\frac{1}{\varepsilon}\right)^2 - 4}\right) \text{ and } k = \sqrt{\frac{\left(1+\frac{1}{\varepsilon}\right)\alpha^2 - 2\alpha}{2\alpha - \left(1+\frac{1}{\varepsilon}\right)}}. \text{ Then:}$$

$$|v| \in \left[\frac{1}{k}|u|,\ k\,|u|\right].$$

The proof was based on the following observation:

Property 4. Let *u* and *v* be non-zero vectors, u_i and v_i be their *i*-th dimensions and $\beta > 0$. Then:

$$\left(u_i - \frac{v_i}{\beta}\right)^2 \geq 0$$

or equivalently,

$$\frac{u_i^{\,2}\beta^2 + v_i^{\,2}}{2\beta} \geq u_i v_i\,.$$

In the proof of Theorem 2, we showed that Property 4 allows us to obtain most strict bounds on lengths of Tanimoto similar vectors for $\beta = \alpha = \frac{1}{2}\left(\left(1+\frac{1}{\varepsilon}\right) + \sqrt{\left(1+\frac{1}{\varepsilon}\right)^2 - 4}\right)$.

Theorem 2 does not take into account the case when $\varepsilon = 1$. However, it is easy to deduce that vectors identical in terms of the Tanimoto similarity have identical lengths.

Theorem 3 [4]. Let u and v be non-zero vectors such that $T(u, v) = 1$. Then:

$$|v| = |u|.$$

In this paper, we will derive bounds on lengths of non-zero real valued vectors that are Tanimoto similar in a simpler way than in [4] and without using Property 4. The new bounds will be also valid for an extended range $(0,1]$ of threshold values of ε.

4 New Bounds on Non-zero Tanimoto Similar Real Valued Vectors

The bounds on lengths of non-zero real valued vectors similar with respect to the Tanimoto similarity we derived in [4] were based on the assumption that $\varepsilon > 1/3$. For smaller value of ε, the proof we provided there would not be valid. In this paper, we derive new bounds and prove them based on Property 5, following from a property of the cosine of the angle of two vectors, instead of Property 4.

Property 5. Let u and v be vectors. Then:

$$u \cdot v \leq |u||v|.$$

Proof. The inequality holds trivially if u or v is a zero vector. Hence, in the remainder of the proof, we consider the case when both u and v are non-zero vectors. Then, $cos(\angle(u,v)) = \dfrac{u \cdot v}{|u||v|}$ and $cos(\angle(u,v)) \leq 1$. So, $u \cdot v = |u||v| cos(\angle(u,v)) \leq |u||v|$. □

Property 6. Let u and v be vectors. Then:

$$|u|^2 + |v|^2 - |u||v| \geq 0.$$

Proof. $|u|^2 + |v|^2 - |u||v| \geq (|u| - |v|)^2 \geq 0$. □

Theorem 4. Let u and v be non-zero vectors such that $T(u, v) \geq \varepsilon$, $\varepsilon \in (0,1]$, $\alpha_1 = \frac{1}{2}\left(\left(1+\frac{1}{\varepsilon}\right) - \sqrt{\left(1+\frac{1}{\varepsilon}\right)^2 - 4}\right)$ and $\alpha_2 = \frac{1}{2}\left(\left(1+\frac{1}{\varepsilon}\right) + \sqrt{\left(1+\frac{1}{\varepsilon}\right)^2 - 4}\right)$. Then:

$$|v| \in [\alpha_1 |u|, \alpha_2 |u|].$$

Proof. By Properties 5 and 6, $T(u,v) = \dfrac{u \cdot v}{|u|^2 + |v|^2 - u \cdot v} \leq \dfrac{|u||v|}{|u|^2 + |v|^2 - |u||v|}$.

Since $\varepsilon \leq T(u,v)$, we may conclude further that $\varepsilon \leq \dfrac{|u||v|}{|u|^2 + |v|^2 - |u||v|}$. In addition, Property 6 allows us to rewrite this inequality as follows:

$\varepsilon|v|^2 - (\varepsilon+1)|u||v| + \varepsilon|u|^2 \leq 0$. Given $\varepsilon \in (0,1)$, we find that the obtained square inequality is fulfilled for $|v| \in [\alpha_1|u|, \alpha_2|u|]$. □

In Property 7, we state a simple property bounding coefficients α_1 and α_2 from Theorem 4, which will allow us to rewrite it as Theorem 5, that uses only one coefficient (α).

Property 7. Let $\varepsilon \in (0,1]$, $\alpha_1 = \frac{1}{2}\left(\left(1+\frac{1}{\varepsilon}\right) - \sqrt{\left(1+\frac{1}{\varepsilon}\right)^2 - 4}\right)$ and $\alpha_2 = \frac{1}{2}\left(\left(1+\frac{1}{\varepsilon}\right) + \sqrt{\left(1+\frac{1}{\varepsilon}\right)^2 - 4}\right)$. Then:

$$\alpha_1 = \frac{1}{\alpha_2}.$$

Proof. The property follows from the fact that $\alpha_1\alpha_2$ = 1; namely, $\alpha_1\alpha_2$ = $\frac{1}{4}\left(\left(1+\frac{1}{\varepsilon}\right)^2 - \left(\left(1+\frac{1}{\varepsilon}\right)^2 - 4\right)\right) = \frac{4}{4} = 1$. □

Theorem 5. Let u and v be non-zero vectors such that $T(u, v) \geq \varepsilon$, $\varepsilon \in (0,1]$ and $\alpha = \frac{1}{2}\left(\left(1+\frac{1}{\varepsilon}\right) + \sqrt{\left(1+\frac{1}{\varepsilon}\right)^2 - 4}\right)$. Then:

$$|v| \in \left[\frac{1}{\alpha}|u|, \alpha|u|\right].$$

Proof. By Theorem 4 and Property 7. □

5 Comparison of Different Bounds on Lengths of Tanimoto Similar Vectors

For comparison purposes, in Table 1, we provide the values of coefficients determining how many times the lengths of Tanimoto similar vectors can be longer/shorter than the length of a given vector, say u, for different values of similarity threshold ε, according to the recalled and newly derived bounds. The first column in Table 1 provides the values of ε. The second column determines how many times the lengths of vectors v with two-valued and/or three-valued dimensions as described in Section 3 in a data set can differ from the length of vector u provided $T(u, v) \geq \varepsilon$ and $\varepsilon \in$ (0, 1]. In particular, it follows from Table 1 that for ε = 0.90 such vectors satisfying condition $T(u, v) \geq \varepsilon$ can be longer than vector u at most 1.05 times and can be shorter than u at most 1.05 times. Columns 3 and 4 provide analogous information for the Tanimoto similarity in the case of real valued vectors. Column 3 provides bounds calculated as specified in [4] in Theorem 2 for $\varepsilon \in \left(\frac{1}{3},1\right)$ and Theorem 3 for $\varepsilon = 1$, while column 4 provides bounds calculated according to Theorem 5 for $\varepsilon \in (0,1]$, as proposed and proved in this paper. One may observe that for the values of ε that are close to 1, only vectors with lengths almost equal to the length of *u* have a chance to be similar to *u*. The smaller the value of ε, the less effective length filtering of vectors that have a chance to be sufficiently similar to vector *u*.

Table 1. Coefficients determining how many times the lengths of Tanimoto similar vectors can be longer/shorter than the length of a given vector

	Theorem 1 (*vectors with two/three-valued domains*)	Theorems 2-3 (*any vectors*)	Theorem 5 (*any vectors*)
ε	$\frac{1}{\sqrt{\varepsilon}}$	k	α
1.0000	1.00	1.00	1.00
0.9999	1.00	1.01	1.01
0.999	1.00	1.03	1.03
0.99	1.01	1.11	1.11
0.98	1.01	1.15	1.15
0.97	1.02	1.19	1.19
0.96	1.02	1.23	1.23
0.95	1.03	1.26	1.26
0.90	1.05	1.39	1.39
0.85	1.08	1.52	1.52
0.80	1.12	1.64	1.64
0.75	1.15	1.77	1.77
0.70	1.20	1.90	1.90
0.65	1.24	2.05	2.05
0.60	1.29	2.22	2.22
0.55	1.35	2.40	2.40
0.50	1.41	2.62	2.62
0.45	1.49	2.87	2.87
0.40	1.58	3.19	3.19
0.35	1.69	3.58	3.58
0.34	1.71	3.67	3.67
0.25	2.00	-	4.79
0.20	2.24	-	5.83
0.15	2.58	-	7.53
0.10	3.16	-	10.91
0.05	4.47	-	20.95

Surprisingly, the bounds provided in columns 3 and 4 are identical for common threshold values; that is, for $\varepsilon \in \left(\frac{1}{3},1\right]$. While the equality of the bounds specified by

Theorem 3 and Theorem 5 for $\varepsilon = 1$ is straightforward, the equivalence of Theorems 2 and 5, would require that $k = \sqrt{\dfrac{\left(1+\frac{1}{\varepsilon}\right)\alpha^2 - 2\alpha}{2\alpha - \left(1+\frac{1}{\varepsilon}\right)}}$ equals α for $\varepsilon \in \left(\frac{1}{3}, 1\right)$. Beneath, we formally prove this equality even for an extended range of ε threshold values; namely, for $\varepsilon \in (0,1)$ (please see Theorem 6).

Theorem 6. Let *u* and *v* be non-zero vectors such that *T*(*u*, *v*) ≥ *ε*, $\varepsilon \in (0,1)$ and $\alpha = \frac{1}{2}\left(\left(1+\frac{1}{\varepsilon}\right) + \sqrt{\left(1+\frac{1}{\varepsilon}\right)^2 - 4}\right)$. Then:

$$\sqrt{\frac{\left(1+\frac{1}{\varepsilon}\right)\alpha^2 - 2\alpha}{2\alpha - \left(1+\frac{1}{\varepsilon}\right)}} = \alpha .$$

Proof (by contradiction). Let $\alpha = \frac{1}{2}\left(\left(1+\frac{1}{\varepsilon}\right) + \sqrt{\left(1+\frac{1}{\varepsilon}\right)^2 - 4}\right)$ and $\sqrt{\dfrac{\left(1+\frac{1}{\varepsilon}\right)\alpha^2 - 2\alpha}{2\alpha - \left(1+\frac{1}{\varepsilon}\right)}} \neq \alpha$. Since $1+\frac{1}{\varepsilon} > 2$ for $\varepsilon \in (0,1)$, then $\alpha > 1$, $\left(1+\frac{1}{\varepsilon}\right)\alpha^2 - 2\alpha > 0$ and $2\alpha - \left(1+\frac{1}{\varepsilon}\right) > 0$. Hence, $\sqrt{\dfrac{\left(1+\frac{1}{\varepsilon}\right)\alpha^2 - 2\alpha}{2\alpha - \left(1+\frac{1}{\varepsilon}\right)}} \neq \alpha$ iff $\left(1+\frac{1}{\varepsilon}\right)\alpha^2 - 2\alpha \neq \alpha^2\left(2\alpha - \left(1+\frac{1}{\varepsilon}\right)\right)$ iff $\alpha^2 - \left(1+\frac{1}{\varepsilon}\right)\alpha + 1 \neq 0$ iff $\alpha \neq \frac{1}{2}\left(\left(1+\frac{1}{\varepsilon}\right) - \sqrt{\left(1+\frac{1}{\varepsilon}\right)^2 - 4}\right)$ and $\alpha \neq \frac{1}{2}\left(\left(1+\frac{1}{\varepsilon}\right) + \sqrt{\left(1+\frac{1}{\varepsilon}\right)^2 - 4}\right)$, which contradicts the assumption. □

6 Experimental Results

In Table 2, we show experimental results we obtained for the data sets: *camcisi*, *classic*, *cranmed*, *hitech*, *sports*, which are available from http://glaros.dtc.umn.edu/gkhome/cluto/cluto/download. For each of these data sets, we provide its characteristics and the percentage of vectors that fulfill the length condition on Tanimoto similarity vectors according to Theorem 5 for different values of ε. If $\varepsilon = 1$, this percentage equals the average number of vectors that have identical length as an individual vector in a data set. Clearly, the percentage of such vectors for $\varepsilon < 1$ cannot be lower than the percentage for $\varepsilon = 1$ for an individual data set. As follows from Table 2, Theorem 5 allows limiting the set of vectors that have a chance to be Tanimoto similar considerably for ε threshold values close to 1.

Table 2. Average percentage of vectors that fulfill the length condition on Tanimoto similarity vectors according to Theorem 5

data set characteristics	data set name	*camcisi*	*classic*	*cranmed*	*hitech*	*sports*
	no. of records	4663	7094	2431	2301	8580
	min. length of a vector	1.00	1.00	4.00	2.65	1.41
	max. length of a vector	30.35	37.22	37.22	125.71	236.41
	avg. length of a vector	4.67	7.40	12.63	25.85	22.15
ε	1.0000	7.33%	3.26%	0.36%	0.12%	0.10%
	0.9999	7.45%	3.63%	1.63%	1.35%	1.09%
	0.999	7.82%	4.66%	5.05%	4.15%	3.42%
	0.99	13.07%	9.67%	15.89%	13.01%	10.79%
	0.95	29.96%	21.85%	35.16%	29.03%	24.17%
	0.90	37.95%	28.86%	49.18%	40.93%	34.32%
	0.85	45.05%	34.55%	59.43%	49.92%	42.23%
	0.80	50.09%	38.84%	67.62%	57.39%	49.01%
	0.75	52.74%	41.76%	74.41%	63.81%	55.05%
	0.70	54.47%	44.04%	80.12%	69.44%	60.56%

7 Conclusions

In this work, we first recalled the bounds on lengths of Tanimoto similar vectors with two-valued and/or three-valued dimensions of a particular type for $\varepsilon \in (0, 1]$ as well as for real valued vectors for $\varepsilon \in (1/3, 1]$, as offered recently in [4]. Next, we derived the new bounds for real valued vectors for the range of ε values extended to (0, 1]. The formulae expressing the new bounds are much simpler than the ones expressing the bounds offered earlier for real valued vectors. Then we analyzed the values of the recalled and new bounds for sample values of ε threshold and found that in the case of real valued vectors the earlier and the new bounds return identical values for common threshold values of ε though the formulae expressing them are quite different and the derivation methods are based on different properties. Eventually, we formally proved that the formulae expressing recalled and new bounds on lengths of Tanimoto similar real valued vectors are equivalent for all values ε in the range (0, 1]. The experimental results we obtained show that these bounds on lengths of similar vectors are a very effective vectors filtering tool for ε close to 1. The filtering of vectors based on these bounds can be easily implemented e.g. by means of a B^+-tree index with the indexing field determined by lengths of vectors.

Acknowledgements. This work was supported by the National Centre for Research and Development (NCBiR) under Grant No. SP/I/1/77065/10 devoted to the Strategic scientific research and experimental development program: "Interdisciplinary System for Interactive Scientific and Scientific-Technical Information".

References

1. Han, J., Kamber, M., Pei, J.: Data Mining: Concepts and Techniques, 3rd edn. Morgan Kaufmann Publishers (2011)
2. Kristensen, T.G.: Transforming Tanimoto Queries on Real Valued Vectors to Range Queries in Euclidian Space. Journal of Mathematical Chemistry 48(2), 287–289 (2010)
3. Kryszkiewicz, M.: Efficient Determination of Binary Non-negative Vector Neighbors with Regard to Cosine Similarity. In: Jiang, H., Ding, W., Ali, M., Wu, X. (eds.) IEA/AIE 2012. LNCS, vol. 7345, pp. 48–57. Springer, Heidelberg (2012)
4. Kryszkiewicz, M.: Bounds on Lengths of Vectors Similar with Regard to the Tanimoto and Cosine Similarity. ICS Research Report 3, Institute of Computer Science, Warsaw University of Technology, Warsaw (2012)
5. Lipkus, A.H.: A proof of the triangle inequality for the Tanimoto dissimilarity. Journal of Mathematical Chemistry 26(1-3), 263–265 (1999)
6. Rajaraman, A., Ullman, J.D.: Mining of Massive Datasets. Cambridge University Press (2011)
7. Salton, G., Wong, A., Yang, C.S.: A vector space model for automatic indexing. Communications of the ACM 18(11), 613–620 (1975)
8. Willett, P., Barnard, J.M., Downs, G.M.: Chemical similarity searching. J. Chem. Inf. Comput. Sci. 38(6), 983–996 (1998)
9. Witten, I.H., Moffat, A., Bell, T.C.: Managing Gigabytes: Compressing and Indexing Documents and Images. Morgan Kaufmann (1999)
10. Zezula, P., Amato, G., Dohnal, V., Batko, M.: Similarity Search: The Metric Space Approach. Springer (2006)

A Quadratic Algorithm for Testing of Z-Codes

Nguyen Dinh Han[1,⋆], Dang Quyet Thang[2], and Phan Trung Huy[3]

[1] Hung Yen University of Technology and Education, Vietnam
hannguyen@utehy.edu.vn

[2] Nam Dinh University of Technology and Education, Vietnam
thangdgqt@gmail.com

[3] Hanoi University of Science and Technology, Vietnam
huyfr2002@yahoo.com

Abstract. We consider a subclass of circular codes, namely Z-codes for bi-infinite words, that have numerous interesting properties appeared in many problems of combinatorics on words. Our major concern is a very basic problem, which is to test whether a language of finite words is a Z-code. As the main result of this paper, we give an efficient algorithm running in a quadratic polynomial time for testing of Z-codes when they are regular.

Keywords: quadratic algorithm, bi-infinite word, monoid, graph, Z-code.

1 Introduction

The bi-infinite words play a crucial role in research infinite behaviors of logical models, formal dynamic systems, games and new code construction, etc. Z-codes for bi-infinite words constitute a subclass of circular codes, that have numerous interesting properties appeared in many problems of combinatorics on words [5,10,1]. The study of Z-codes in formal languages, specially in theory of codes has attracted many works [7,8,4,11,9,3,5,10,2,1], which showed the important role of Z-codes.

Testing whether a language of finite words is a Z-code is a fundamental problem in theory of codes. The testing algorithm for Z-codes for the case of finite languages is given in [9]. However, for the general case of regular languages, such an algorithm is not known and it is the subject of this paper. Here we propose a new test and establish a quadratic testing algorithm with time complexity $\mathcal{O}(n^2)$ to verify if a regular language X saturated by a monoid morphism is a Z-code, where n is the index of the syntactic congruence of X.

The content of this paper is organized as follows. In Section 2, we recall some basic notions of languages, codes of finite, infinite and bi-infinite words. A new technique based on finite monoids and graphs to establish a test for Z-codes

[⋆] This work is partially supported by Vietnamese National Foundation for Science and Technology Development (NAFOSTED).

A. Selamat et al. (Eds.): ACIIDS 2013, Part I, LNAI 7802, pp. 455–464, 2013.

is presented. As a consequence, we obtain an effective testing algorithm for Z-codes with time complexity $\mathcal{O}(n^2)$, which is the main result of the paper and it is presented in Section 3.

2 Preliminaries

In the following, we recall some notions (for more details, we refer to [1,6,11]). Let A be a finite alphabet. As usual, A^* is the free monoid generated by A. The empty word is denoted by ε and $A^+ = A^* - \{\varepsilon\}$. The length $|w|$ of the word $w = a_1a_2 \cdots a_n$ with $a_i \in A$ is n. By convention $|\varepsilon| = 0$. A subset of A^* is called a *language*. For any language $X \subseteq A^*$, we denote by X^* the submonoid of A^* generated by X, $X^* = X^+ \cup \{\varepsilon\}$. A *factorization* of a word w in $X^+$is a finite sequence $\{u_1, u_2, \ldots, u_n\}$ of words of X such that $w = u_1u_2 \cdots u_n$, $n \geq 1$. A language $X \subseteq A^+$ is a *code* if every word w in A^* has at most one factorization on X.

A *right infinite* (resp. *left infinite, bi-infinite*) ω-word is a sequence $a_1a_2 \cdots$ (resp. $\cdots a_2a_1$, $\cdots a_{-2}a_{-1}a_0a_1a_2 \cdots$), $a_i \in A$. Let ${}^N\!A, A^N, A^Z$ be sets of left infinite, right infinite and bi-infinite ω-words on A, respectively. For each language X of A^*, we denote by ${}^\omega X, X^\omega$ and ${}^\omega X^\omega$ left infinite, right infinite and bi-infinite product of non-empty words in X. We call each subset of A^N (resp. ${}^N\!A, A^Z$) an *ω-language* of right infinite (resp. left infinite, bi-infinite) ω-words. A *factorization* of a word $\alpha \in X^\omega$ (resp. $\alpha \in {}^\omega X$) is an infinite sequence of words $\{u_1, u_2, \ldots\}$ in X such that $\alpha = u_1u_2 \cdots$ (resp. $\alpha = \cdots u_2u_1$). Let $w \in A^Z$ be in the form: $w = \cdots a_{-2}a_{-1}a_0a_1a_2 \cdots$, with $a_i \in A$. A *factorization* on elements of X of the bi-infinite ω-word w is a strictly increasing function $\mu : Z \to Z$ satisfying $x_i = a_{\mu(i)+1} \cdots a_{\mu(i+1)} \in X$ for all $i \in Z$.

Two factorizations μ and λ are said to be *equal*, denoted $\mu = \lambda$ if there is $t \in Z$ such that $\lambda(i + t) = \mu(i)$ for all $i \in Z$. Otherwise, λ and μ are *distinct*, denoted $\mu \neq \lambda$. It is easy to verify that $\mu \neq \lambda$ if and only if $\mu(Z) \neq \lambda(Z)$, or equivalently, there exist a word $u \in A^+$, two bi-infinite sequences of words of X: $\ldots, x_{-2}, x_{-1}, x_0, x_1, x_2, \ldots$ and $\ldots, y_{-2}, y_{-1}, y_0, y_1, y_2, \ldots$ such that

$$\cdots x_{-2}x_{-1}u = \cdots y_{-1}y_0,\ |u| \leq |x_0|,$$

$$x_0x_1 \cdots = uy_1y_2 \cdots,\ |u| \leq |y_0| \text{ with } u \neq x_0 \text{ or } u \neq y_0.$$

A language X of A^+ is a *Z-code* if all factorizations on X of every bi-infinite word are equal.

Let M be a monoid. For $S, T \subseteq M$, we define *left quotients* and *right quotients* of S by T as follows: $T^{-1}S = \{u \in M \mid \exists t \in T, t.u \in S\}$, $ST^{-1} = \{u \in M \mid \exists t \in t, u.t \in S\}$. The notations $u^{-1}S, Su^{-1}$ will be used when $T = \{u\}$ is singleton. For any $u, v \in M$, we write uv instead of $u.v$ whenever $M = A^*$.

Given $X \subseteq A^*$, we say that X *is saturated by a monoid morphism* $\varphi : A^* \to M$ if there exists $B \subseteq M$ such that $X = \varphi^{-1}(B)$ and in that case, we also say that M saturates X, and X is given by this tuple (φ, M, B). In case X is regular, M can be chosen by the transition monoid of the minimal automaton recognizing X, or by the syntactic monoid M_X. We say that $k = \mathrm{Card}(M_X)$ *is the index of the syntactic congruence of* X, briefly k *is the index of* X.

3 Testing for Rational Z-Codes

3.1 A Language Algorithm

Let $X \subseteq A^+$ be a regular language. We define the set of *overlap* of elements of X by

$W(X) = ((A^+)^{-1}X \cap X) \cup (X(A^+)^{-1} \cap X) \cup (((A^+)^{-1}X \cap X(A^+)^{-1}) - \{\varepsilon\})$
Let $U = W(X)$, we consider two sequences of regular sets U_i, V_i defined recursively as follows

$$\begin{aligned} U_1 &= (UX^*)^{-1}X - \{\varepsilon\}, U_{i+1} = (U_iX^*)^{-1}X, \\ V_1 &= X(UX^*)^{-1} - \{\varepsilon\}, V_{i+1} = X(V_iX^*)^{-1} \end{aligned} \tag{1}$$

for all $i \geq 1$.

Definition 1. *Let $X \subseteq A^+$ and let $u \in A^+$. A right infinite (resp. left infinite) ω-word α is said to be right infinite (resp. left infinite) u-ambiguous on X if there exist $x_i, y_j \in X$, $i, j = 1, 2, \ldots$ such that $\alpha = x_1x_2\cdots = uy_1y_2\cdots$ with $|u| < |x_1|$ (resp. $\alpha = \cdots x_2x_1u = \cdots y_2y_1$ with $|u| < |y_1|$). Whenever X is defined, for simplicity, we call α right infinite (resp. left infinite) u-ambiguous.*

We establish the following results for the correctness of the test for Z-codes.

Lemma 1. *Let $X \subseteq A^+$ be a regular language, $U = W(X)$ and let $U_i, V_i\, (i \geq 1)$ be defined in (1). Then, for every $u \in U$,*
(i) there is no right infinite u-ambiguous ω-word if and only if there exists $i \geq 1$ such that $U_i = \emptyset$.
(ii) there is no left infinite u-ambiguous ω-word if and only if there exists $i \geq 1$ such that $V_i = \emptyset$.

Proof. (i) $(\Rightarrow)$. We assume by a contradiction that $U_i \neq \emptyset$ for all $i \geq 1$. Let N be any integer not less than the index n of X, and let $u_N \in U_N$. Then, there exist $u_i \in U_i$, $i = 1, 2, \ldots, N-1$ such that $u_1 \in (uX^*)^{-1}X, u_{i+1} \in (u_iX^*)^{-1}X, i = 1, 2, \ldots, N-1$ with $\varepsilon \neq u \in U$, or equivalently, $uu_1 \in X$, $u_iX^*u_{i+1} \in X$, $i = 1, 2, \ldots, N-1$. Among $u_1, u_2, \ldots, u_N$ we can pick out u_q and u_p such that $p < q$ and $u_p \sim_X u_q$. We define now an infinite sequence of words $u'_1, u'_2, \ldots$ by putting $u'_i = u_i$, $1 \leq i \leq q-1$ and $u'_{q+i} = u_{p+t}$, $0 \leq i \leq q-1$, where t is the least nonnegative residue of i mod $(q-p)$. It is easy to verify that $x'_i = u'_iX^*u'_{i+1} \in X^*$ for $i = 1, 2, \ldots$ and $x' = uu'_1 = uu_1 \in X$.

Consider now the infinite product of words $uu'_1X^*u'_2X^*\cdots$ written in two ways: $(uu'_1)(X^*)(u'_2X^*u'_3)\cdots = u(u'_1X^*u'_2)(u'_3X^*u'_4)\cdots$ or $x_1x_2\cdots = uy_1y_2\cdots$ with $x_1 \in X$, $|u| < |x_1|$, $x_i, y_j \in X$. Thus, the right infinite ω-word $\alpha = x_1x_2\cdots = uy_1y_2\cdots$ with $|u| < |x_1|$ is right infinite u-ambiguous. This contradicts assumption.

$(\Leftarrow)$. We assume by a contradiction that there exist $u \in U$ and $x_i, y_j \in X$, $i, j = 1, 2, \ldots$ such that $x_1x_2\cdots = uy_1y_2\cdots$ with $|u| < |x_1|$. The proof is proceeded if the following assertion can be verified: For all $k \geq 1$, there exist a word $z \in U_k$ and two integers $i \geq 1, j \geq 0$ such that one of the following two conditions holds

(a)
$$x_1 \cdots x_i z = u y_1 \cdots y_j$$
$$x_{i+1}x_{i+2} \cdots = z y_{j+1} y_{j+2} \cdots \text{ with } |u| < |x_1| \text{ and } |z| \leq |y_j|.$$
(b)
$$x_1 \cdots x_i = u y_1 \cdots y_j z$$
$$z x_{i+1}x_{i+2} \cdots = y_{j+1} y_{j+2} \cdots \text{ with } |u| < |x_1| \text{ and } |z| \leq |x_i|.$$

The proof is by an induction on k. From $|u| < |x_1|$ we have $z \in A^+$ such that $x_1 = uz$. Then $z = u^{-1}x_1 \in (UX^*)^{-1}X - \{\varepsilon\} = U_1$. Thus, we have
$$x_1 = uz$$
$$z x_2 x_3 \cdots = y_1 y_2 \cdots \text{ with } |u| < |x_1| \text{ and } |z| < |x_1|.$$

Therefore the assertion holds for $k = 1$. Assume now that the assertion holds for some $k > 1$, we prove that it remains true for $k + 1$. By assumption, there exist a word $z \in U_k$ and two integer $i \geq 1, j \geq 0$ such that (a) or (b) holds. Suppose (a) holds, then we have
$$x_1 x_2 \cdots x_i z = u y_1 \cdots y_j$$
$$x_{i+1}x_{i+2} \cdots = z y_{j+1} y_{j+2} \cdots \text{ with } |u| < |x_1| \text{ and } |z| \leq |y_j|.$$

We distinguish two cases, according to the length of z compared to the length of x_{i+1},

Case 1: $|z| \leq |x_{i+1}|$. We have $z' \in A^*$ such that $x_{i+1} = zz'$. Thus $z' = z^{-1}x_{i+1} \in (U_k X^*)^{-1}X = U_{k+1}$. We have
$$x_1 \cdots x_{i+1} = u y_1 \cdots y_j z'$$
$$z' x_{i+2}x_{i+3} \cdots = y_{j+1} y_{j+2} \cdots \text{ with } |u| < |x_1| \text{ and } |z'| \leq |x_{i+1}|.$$

This implies that (b) holds.

Case 2: $|z| > |x_{i+1}|$. There exist $p \geq i + 1$ and $z' \in A^*$ such that $z = x_{i+1}x_{i+2} \cdots x_p z'$ and $|z'| \leq |x_{p+1}|$. Moreover, from $|z| \leq |y_j|$ we have $u' \in A^*$ such that $y_j = u'z$, or equivalently, $y_j = u'x_{i+1}x_{i+2} \cdots x_p z'$. Since $z = u'^{-1}y_j \in U_k = (U_{k-1}X^*)^{-1}X$, we have $u' \in U_{k-1}X^*$. Thus $u'x_{i+1}x_{i+2} \cdots x_p \in U_{k-1}X^*$. It implies that $z' = (u'x_{i+1}x_{i+2} \cdots x_p)^{-1}y_j \in (U_{k-1}X^*)^{-1}X = U_k$. Then, there exists $z'' \in A^*$ such that $|z''| \leq |x_{p+1}|$ and $z'' = z'^{-1}x_{p+1} \in (U_k X^*)^{-1}X = U_{k+1}$. Hence
$$x_1 \cdots x_{p+1} = u y_1 \cdots y_j z''$$
$$z'' x_{p+2}x_{p+3} \cdots = y_{j+1} y_{j+2} \cdots \text{ with } |u| < |x_1| \text{ and } |z''| \leq |x_{p+1}|.$$

This implies that (b) holds. Suppose now (b) holds, then we have
$$x_1 \cdots x_i = u y_1 \cdots y_j z$$
$$z x_{i+1}x_{i+2} \cdots = y_{j+1} y_{j+2} \cdots \text{ with } |u| < |x_1| \text{ and } |z| \leq |x_i|.$$

Compare the length of z and the length of y_{j+1}, we consider two cases similarly to the above proof and our conclusion is that (a) holds for these two cases.

Therefore the assertion holds for all $k \geq 1$. It implies that $U_i \neq \emptyset$ for all $i \geq 1$. This contradicts the assumption.

The proof of (i) is completed. The proof of (ii) is proceeded similarly to the proof of (i). □

Now, we can formulate the main result of this section in the following theorem, which gives a criterion for Z-codes.

Theorem 1. *Let $X \subseteq A^+$ be a regular language, $U = W(X)$ and let U_i, V_i $(i \geq 1)$ be defined in* (1)*. Then, X is a Z-code if and only if for every $u \in U$,*

(i) if $u \in U \cap X$ then there is no right infinite u-ambiguous ω-word and there is no left infinite u-ambiguous ω-word;
(ii) if $u \in U \cap X$ then there is no right infinite u-ambiguous ω-word or there is no left infinite u-ambiguous ω-word.

Proof. $(\Rightarrow)$. Assume that X is not a Z-code and assume (i) or (ii) does not hold. Suppose (i) does not hold. Then, there exist $u \in U \cap X$ and $x_i, y_j \in X$, $i, j = -2, -1, 0, 1, 2, \ldots$ such that $x_0x_1 \cdots = uy_1y_2 \cdots$ with $|u| < |x_0|$, or $\cdots x_{-2}x_{-1}u = \cdots y_{-2}y_{-1}y_0$ with $|u| < |y_0|$. If $x_0x_1 \cdots = uy_1y_2 \cdots$ with $|u| < |x_0|$, then there exist two bi-infinite sequences of words in X: $\ldots, x_{-2}, x_{-1}, x_0, x_1, x_2, \ldots$ and $\ldots, y_{-2} = x_{-2}, y_{-1} = x_{-1}, y_0 = u, y_1, y_2, \ldots$ such that

$$\cdots x_{-2}x_{-1}u = \cdots y_{-2}y_{-1}y_0,\ |u| < |x_0|,$$
$$x_0x_1 \cdots = uy_1y_2 \cdots,\ |u| = |y_0| \text{ with } u \neq x_0.$$

Thus X is not a Z-code. This is a contradiction.

If $\cdots x_{-2}x_{-1}u = \cdots y_{-2}y_{-1}y_0$ with $|u| < |y_0|$, then there exist two bi-infinite sequences of words in X: $\ldots, x_{-2}, x_{-1}, x_0 = u, x_1, x_2, \ldots$ and $\ldots, y_{-2}, y_{-1}, y_0$, $y_1 = x_1, y_2 = x_2, \ldots$ such that

$$\cdots x_{-2}x_{-1}u = \cdots x_{-2}x_{-1}y_0,\ |u| = |x_0|,$$
$$x_0x_1 \cdots = uy_1y_2 \cdots,\ |u| < |y_0| \text{ with } u \neq y_0.$$

Thus X is not a Z-code. This contradicts with the assumption.

Conversely, suppose that (ii) does not hold, then there exist $u \in U \cap X$ and $x_i, y_j \in X, i, j = -2, -1, 0, 1, 2, \ldots$ such that $x_0x_1 \cdots = uy_1y_2 \cdots$ with $|u| < |x_0|$, and $\cdots x_{-2}x_{-1}u = \cdots y_{-2}y_{-1}y_0$ with $|u| < |y_0|$. Then, by definition, X is not a Z-code. This is a contradiction.

$(\Leftarrow)$. We assume by a contradiction that X is not a Z-code. Then, there exist relations

$$\cdots x_{-2}x_{-1}u = \cdots y_{-2}y_{-1}y_0,\ |u| \leq |x_0|,$$
$$x_0x_1 \cdots = uy_1y_2 \cdots,\ |u| \leq |y_0| \text{ with } u \neq x_0 \text{ or } u \neq y_0.$$

From $|u| \leq |x_0|$ and $|u| \leq |y_0|$, we have $u \in U$. We distinguish two cases

Case 1: $u \in U \cap X$. Then, if $u \neq x_0$, we have $x_0x_1 \cdots = uy_1y_2 \cdots$, $|u| < |x_0|$. Conversely, if $u \neq y_0$, we have $\cdots x_{-2}x_{-1}u = \cdots y_{-2}y_{-1}y_0$, $|u| < |y_0|$. It implies that (i) does not hold. This contradicts with the assumption.

Case 2: $u \in U - X$. Then, $u \neq x_0$ and $u \neq y_0$. We have $x_0x_1 \cdots = uy_1y_2 \cdots$, $|u| < |x_0|$, and $\cdots x_{-2}x_{-1}u = \cdots y_{-2}y_{-1}y_0$, $|u| < |y_0|$. It implies that (ii) does not hold. This is a contradiction.

The proof is completed. □

Remark 1. For finite languages, Theorem 1 provides us a testing algorithm for Z-codes. But in general case, to establish testing algorithms for Z-codes when they are regular languages, we have to associate with them finite structures such as finite monoids or finite graphs. This will be demonstrated in the next sections.

3.2 An Algorithm Basing on Monoids

Let $h : A^* \to P$ be a surjective monoid morphism saturating both $Y = \{\varepsilon\}$ and $X \subseteq A^+$, where P is a finite monoid, $X = h^{-1}(K)$, $Y = h^{-1}(L)$, $L = \{1_P\}, Q = h(A^+)$ with $K, L, Q \subseteq P$. Consider the set $\mho(K)$ of elements of K is defined by:

$$\mho(K) = (Q^{-1}K \cap K) \cup (KQ^{-1} \cap K) \cup ((Q^{-1}K \cap KQ^{-1}) - L).$$

By definition, one can verify the following basic results

Lemma 2. *Let $X \subseteq A^+$ be a regular language and let $h : A^* \to P$ be a surjective monoid morphism saturating both $Y = \{\varepsilon\}$ and X, where P is a finite monoid, $X = h^{-1}(K)$, $Y = h^{-1}(L)$, $L = \{1_M\}, Q = h(A^+)$, $K, L, Q \subseteq P$. Then,* $W(X) = h^{-1}(\mho(K))$.

Definition 2. *Let $X \subseteq A^+$ be a regular language and let $h : A^* \to P$ be a surjective monoid morphism saturating both $Y = \{\varepsilon\}$ and X, where P is a finite monoid, $X = h^{-1}(K)$, $Y = h^{-1}(L)$, $L = \{1_M\}, Q = h(A^+)$, $K, L, Q \subseteq P$. Let $O = \mho(K)$. We define*
(i) A $\rho-$cycle of type 1 in P is a sequence $e_1, \ldots, e_j$ such that $e_1, \ldots, e_{j-1}$ are all distinct, but $e_j = e_i$, $j > i$ where $e_1 \in (O.S)^{-1}K - L, e_{i+1} \in (e_i.S)^{-1}K$, $i \geq 1$.
(ii) A $\rho-$cycle of type 2 in P is a sequence $f_1, \ldots, f_j$ such that $f_1, \ldots, f_{j-1}$ are all distinct, but $f_j = f_i$, $j > i$ where $f_1 \in K(O.S)^{-1} - L, f_{i+1} \in K(e_i.S)^{-1}$, $i \geq 1$.
(iii) A $\rho-$cycle of type 1 in P is said to start at $o \in O$ if $e_1 \in (o.S)^{-1}K - L$.
(iv) A $\rho-$cycle of type 2 in P is said to start at $o \in O$ if $f_1 \in K(o.S)^{-1} - L$.

Lemma 3. *Let $X \subseteq A^+$ be a regular language, $U_i, V_i\,(i \geq 1)$ be defined in (1). Let $h : A^* \to P$ be a surjective monoid morphism saturating both $Y = \{\varepsilon\}$ and X, where P is a finite monoid, $X = h^{-1}(K)$, $Y = h^{-1}(L)$, $L = \{1_M\}, S = K^*$, $K, L \subseteq P$ and let $e_1, e_2, \ldots, e_j = e_i, j > i$, $f_1, f_2, \ldots, f_j = f_i, j > i$ be ρ-cycles of type 1 and 2 in P. Then, we have $h^{-1}(e_i) \subseteq U_i$ and $h^{-1}(f_i) \subseteq V_i$ for all $i \geq 1$.*

Lemma 4. *Let $X \subseteq A^+$ be a regular language and let $h : A^* \to P$ be a surjective monoid morphism saturating both $Y = \{\varepsilon\}$ and X, where P is a finite monoid, $X = h^{-1}(K)$, $Y = h^{-1}(L)$, $L = \{1_M\}, S = K^*$, $K, L \subseteq P$. Let $U = W(X)$. Then,*
(i) there exists a right infinite u-ambiguous ω-word with $u \in U$ if and only if there exists a $\rho-$cycle of type 1 in P.
(ii) there exists a left infinite u-ambiguous ω-word with $u \in U$ if and only if there exists a $\rho-$cycle of type 2 in P.

Proof. (i) $(\Rightarrow)$. Indeed, if there exists a right infinite u-ambiguous ω-word with $u \in U$, then by Lemma 1, $U_i \neq \emptyset$ for all $i \geq 1$. Let $N = \text{Card}(P) + 1$, the we have a sequence $u_1, u_2, \ldots, u_{N-1}, u_N$ with $u_1 \in (uX^*)^{-1}X - \{\varepsilon\}$, $u_{i+1} \in (u_iX^*)^{-1}X$, $N > i \geq 1$. From the sequence $u_1, u_2, \ldots, u_{N-1}, u_N$, we show that there exists a $\rho-$cycle of type 1 in P. At first, let $e_1 = h(u_1)$. Then $e_1 \neq 1_P$ whence $u_1 \neq \varepsilon$. We have: $u_1 \in (uX^*)^{-1}X - \{\varepsilon\} \Leftrightarrow (\exists y \in X^*, x \in X : \varepsilon \neq u_1 = (uy)^{-1}x \Leftrightarrow x = uyu_1, u_1 \neq \varepsilon)$. It follows from $y \in X^*$ that there exist $\lambda \geq 0$ and $x_1, x_2, \ldots, x_\lambda \in X$ such that $y = x_1.x_2.\cdots.x_\lambda$. Then $x = uyu_1 = ux_1x_2\ldots x_\lambda u_1$ and $u_1 \neq \varepsilon$. We have $h(x) = h(u).h(x_1).h(x_2).\ldots.h(x_\lambda).h(u_1)$ with $h(x), h(x_1), h(x_2), \ldots, h(x_\lambda) \in K, h(u) \in O = \mho(K)$ and $h(u_1) \neq 1_P$. Thus $h(u_1) = e_1 = (h(u).h(x_1).h(x_2).\ldots.h(x_\lambda))^{-1}h(x) \in (O.S)^{-1}K - L$. Next, let $e_2 = h(u_2)$, we have: $u_2 \in (u_1X^*)^{-1}X \Leftrightarrow (\exists y = x_1x_2\cdots x_\lambda \in X^*, x \in X : u_2 = (u_1y)^{-1}x \Leftrightarrow x = u_1x_1x_2\cdots x_\lambda u_2)$. Then $h(x) =$

$h(u_1).h(x_1).h(x_2).\cdots.h(x_\lambda).h(u_2)$ with $h(x), h(x_1), h(x_2), \ldots, h(x_\lambda) \in K$. We have: $e_2 = h(u_2) = (h(u_1).h(x_1).h(x_2).\ \cdots.h(x_\lambda))^{-1}h(x) \in (e_1.S)^{-1}K$. Similarly, we have $e_3 \in (e_2.S)^{-1}K, \ldots, e_N \in (e_{N-1}.S)^{-1}K$. If $e_1, e_2, \ldots, e_N$ are all distinct, then the number of different e_i exceeds the cardinality of P. Thus, there must exists $i < j \leq N$ such that $e_j = e_i$. It implies that there exists a $\rho-$cycle of type 1 $e_1, e_2, \ldots, e_j = e_i, j > i$ in P.

($\Leftarrow$). Suppose that there exists a $\rho-$cycle of type 1 $e_1, e_2, \ldots, e_j = e_i, j > i$ in P. Now we put an arrow $e_i \rightarrow e_{i+1}$ if $e_{i+1} \in (e_i.S)^{-1}K$. Then, by assumption, we have a path: $e_1 \rightarrow e_2 \rightarrow \cdots \rightarrow e_i \rightarrow e_{i+1} \rightarrow \cdots \rightarrow e_{j-1} \rightarrow e_j$. We have $e_{i+1} \in (e_i.S)^{-1}K = (e_j.S)^{-1}K$ since $e_j = e_i$. Thus, we have a path

$$e_1 \rightarrow e_2 \rightarrow \cdots \rightarrow e_i \rightarrow e_{i+1} \rightarrow \cdots \rightarrow e_{j-1} \rightarrow e_j \rightarrow e_{i+1}. \tag{2}$$

It implies that there exists an infinite path

$$e_1 \rightarrow e_2 \rightarrow \cdots \rightarrow e_i \rightarrow e_{i+1} \rightarrow \cdots \rightarrow e_{j-1} \rightarrow e_j \rightarrow e_{i+1} \rightarrow e_{i+2} \rightarrow \cdots$$

Then, there exists an infinite sequence $e_1, e_2, \ldots, e_i, e_{i+1}, \ldots, e_j = e_i, e_{j+1} = e_{i+1}, \ldots$, by Lemma 3, $\emptyset \neq h^{-1}(e_i) \subseteq U_i$, $\forall i \geq 1$. This implies that $U_i \neq \emptyset$ for all $i \geq 1$. By Lemma 1, there exists a right infinite u-ambiguous ω-word with $u \in U$. The proof of (i) is completed. The proof of (ii) is proceeded similarly to (i). □

Now, as a direct consequence of Theorem 1, Lemma 1 and Lemma 3, we establish the main result of this section which provides testing algorithms for Z-codes as follows.

Theorem 2. *Let $X \subseteq A^+$ be a regular language and let $h : A^* \rightarrow P$ be a surjective monoid morphism saturating both $Y = \{\varepsilon\}$ and X, where P is a finite monoid, $X = h^{-1}(K)$, $Y = h^{-1}(L)$, $L = \{1_M\}, S = K^*$, $K, L \subseteq P$. Let $O = \mho(K)$. Then, X is not a Z-code if and only if there exists $o \in O$ such that one of the following conditions holds*
(i) if $o \in O \cap K$ then there exists a $\rho-$cycle of type 1 *or a $\rho-$cycle of type* 2 *in P, starting at the same o.*
(ii) if $o \in O - K$ then there exists a $\rho-$cycle of type 1 *and a $\rho-$cycle of type* 2 *in P, starting at the same o.*

3.3 An Algorithm Basing on Graphs

In this section, based on finite graphs, we establish a testing algorithm for Z-codes and calculate its time complexity. From P and $K, S \subseteq P$ defined in Section 3.2, we construct a coloured directed graph $G = (V, E)$ as follows. Initially, set $E = \emptyset$. Set $V = P$ and E is updated as follows.

If $m \in e.S$, then we add a red arc $e \xrightarrow{red} m$ into E.

If $a.b \in K$, or equivalently, $b \in a^{-1}K$ and $a \in Kb^{-1}$ then we add a blue arc $a \xrightarrow{blue} b$ and a yellow arc $b \xrightarrow{yellow} a$ into E.

We have

$$e \xrightarrow{red} m \xrightarrow{blue} e' \Leftrightarrow e' \in (e.S)^{-1}K \tag{3}$$

and

$$f \xrightarrow{red} m \xrightarrow{yellow} f' \Leftrightarrow f' \in K(f.S)^{-1} \tag{4}$$

Then, each path mentioned in (2)

$$e_1 \to e_2 \to \cdots e_i \to e_{i+1} \to \cdots \to e_{j-1} \to e_j \to e_{i+1}$$

and similarly, each path

$$f_1 \to f_2 \to \cdots f_i \to f_{i+1} \to \cdots \to f_{j-1} \to f_j \to f_{i+1}$$

can be extended to the following paths in G

$$e_1 \xrightarrow{red} m_1 \xrightarrow{blue} e_2 \xrightarrow{red} \cdots \xrightarrow{blue} e_{i+1} \xrightarrow{red} \cdots \xrightarrow{blue} e_j \xrightarrow{red} m_j \xrightarrow{blue} e_{i+1} \tag{5}$$

and

$$f_1 \xrightarrow{red} m_1 \xrightarrow{yellow} f_2 \xrightarrow{red} \cdots \xrightarrow{yellow} f_{i+1} \xrightarrow{red} \cdots \xrightarrow{yellow} f_j \xrightarrow{red} m_j \xrightarrow{yellow} f_{i+1} \tag{6}$$

We call the paths (5) and (6) *$\rho-$cycles of type* 1 and *type* 2 in G, respectively. We denote by S_1, S_2 subsets of V, $S_1 = \mho(K), S_2 = K$. Then, a $\rho-$cycle of type 1 (resp. type 2) in G defined above is said to start at a vertex $v \in S_1$ if $e_1 \in (v.S)^{-1}K - L$ (resp. $f_1 \in K(v.S)^{-1} - L$) with $K, S, L \subseteq P$ defined in Section 3.2.

Now, as a direct consequence of Theorem 2, we establish the main result of this section in the following theorem, that provides an effective testing algorithm for Z-codes when they are regular.

Theorem 3. *Let $X \subseteq A^+$ be a regular language and let $G = (V, E)$ be a directed graph with $S_1, S_2 \subseteq V$ be defined above. Then, X is not a Z-code if and only if there exists $v \in S_1$ such that one of the following conditions holds*
(i) if $v \in S_1 \cap S_2$, then there exists a $\rho-$cycle of type 1 *or a $\rho-$cycle of type* 2 *in G, starting at the same vertex v.*
(ii) if $v \in S_1 - S_2$, then there exist a $\rho-$cycle of type 1 *and a $\rho-$cycle of type* 2 *in G, starting at the same vertex v.*

Next, we present the testing algorithm for Z-codes deduced from Theorem 3. Assume that the cardinality of A is a constant, input to the algorithm is a tuple (φ, M, B), where $\varphi : A^* \to M$ is a monoid morphism saturating X, M is a finite monoid, $B \subseteq M$, $X = \varphi^{-1}(B)$ and output of the algorithm is the answer whether X is a Z-code. The algorithm can be established as follows.

Remark 2. By the method introduced in [6], from a given monoid morphism $h : A^* \to M$ saturating a regular language $X \subseteq A^*$, $\mathrm{Card}(M) = n$, define $P = K^*$ with $K = h(A)$ requires a time complexity $\mathcal{O}(n^2)$. Using this P, we can define a surjective monoid morphism $h' : A^* \to P$ which saturates X. Calculating Q, K^+ and K^* from K can be completed in $\mathcal{O}(n^2)$ time.

Theorem 4. *Let $X \subseteq A^+$ be a regular language and let $h : A^* \to P$ be a surjective monoid morphism saturating X. Then, there is an $\mathcal{O}(n^2)$ algorithm to verify if P has some ρ-cycle of type* 1 *and type* 2.

Proof. Indeed, we can assume X is given by a tuple (h, K, P) with $h^{-1}(K) = X$. *Step* 1 (Construct G). From P, K, we can construct $S = K^*, T = K^+$ with the time complexity $\mathcal{O}(n^2)$ [6]. Using S, T, K, P, we construct a graph $G = (V, E)$, with $V = P \times I$, $S_1 = \mho(K), S_2 = K, S_1, S_2 \subseteq V$. Where $I = \{1, 2, 3, 4\}$ and E is updated by the following procedure

Procedure 1. *Contruct G*

```
for $e \in P$ do
  for $s \in P$ do
     if $s \in S$ (or equivalently $flagS(s) == 1$) then
        add $((e, 1) \xrightarrow{red} (e.s, 3))$ and $((e, 2) \xrightarrow{red} (e.s, 4))$ to $E$
for $m \in P$ do
  for $e \in P$ do
     if $m.e \in K$ (or equivalently $flagK(m.e) == 1$) then
        add $((m, 3) \xrightarrow{blue} (e, 1))$ and $((m, 4) \xrightarrow{yellow} (e, 2))$ to $E$
```

Thus, constructing G requires a time complexity $\mathcal{O}(n^2)$.

Step 2 (verify ρ-cycles of type 1 and 2 in G). We use the method introduced in [6] to design a new algorithm which allow us to check whether G has some ρ-cycles of type 1 and 2 in $\mathcal{O}(\text{Card}(V) + \text{Card}(E))$ time. Thus, Step 2 can be completed in $\mathcal{O}(n^2)$ time. The proof is completed. □

Algorithm 1. *Testing algorithm for Z-codes in regular case*

```
Input:   A regular language $X \subseteq A^+$ is given by a tuple $(\varphi, M, B)$
Output:  "YES" if $X$ is a $Z$-code, "NO" otherwise.
Step 1:  From $(\varphi, M, B)$, we construct a surjective monoid morphism
         $h : A^* \to P \subseteq M^1$ saturates both $X$ and $Y = \{\varepsilon\}$ {see Example 1 [6]}
Step 2:  Calculate $K = P \cap B$ that satisfies $h^{-1}(K) = X$
Step 3:  Calculate $Q, S = K^*, T = K^+$
Step 4:  From $P, K, S, T$ construct $G$ with the set of vertices $P \times I = \{1, 2, 3, 4\}$
Step 5:  (Verify $\rho$-cycles of type 1 and 2 in $G$)
         if there exists a $\rho$-cycle of type 1 starting at $v \in S_1 \cap S_2$ then
             return "NO"
         if there exists a $\rho$-cycle of type 2 starting at $v \in S_1 \cap S_2$ then
             return "NO"
         if there exists $\rho$-cycles of type 1 and 2 starting at $o \in S_1 - S_2$ then
             return "NO"
         else return "YES"
```

Details and complexity of the algorithm

1) In Step 1, the time complexity is $\mathcal{O}(n^2)$ to define two arrays $flagP, flagB$ (Lemma 4 [6]).
2) In Step 2, calculating K requires a time complexity $\mathcal{O}(n)$ by conditions: $flagK(x) = 1$ if and only if $flagP(x) = 1$ and $flagB(x) = 1$ for any $x \in M^1$.

3) In Step 3, calculating $T = K^+$ and $S = K^*$ as subsets of P require a time complexity $\mathcal{O}(n^2)$ (Lemma 4 [6]).
4) In Step 4 and Step 5, applying Theorem 4, the time complexity for these steps is $\mathcal{O}(n^2)$.

Therefore, in total we have at most $\mathcal{O}(n^2)$ time complexity for the whole algorithm, where $n = \text{Card}(M)$.

4 Conclusion

In the paper we give an affirmative answer for the question about existence of an effective testing algorithm running in a quadratic polynomial time for testing of Z-codes. Indeed, we establish a quadratic algorithm that, given as input a regular language X defined by a tuple (φ, M, B), where $\varphi : A^* \to M$ is a monoid morphism saturating X, M is a finite monoid, $B \subseteq M$, $X = \varphi^{-1}(B)$, decides in $\mathcal{O}(n^2)$ time whether X is a Z-code, where n is the index of X.

Acknowledgment. This work is partially supported by NAFOSTED. The authors acknowledge their great indebtedness to the referees for their valuable comments and suggestions.

References

1. Berstel, J., Perrin, D., Reutenauer, C.: Codes and automata. Cambridge University Press (2010)
2. Carton, O., Perrin, D., Éric Pin, J.: Automata and semigroups recognizing infinite words. Logic and Automata 42, 133–168 (2008)
3. Devolder, J., Latteux, M., Litovsky, I., Staiger, L.: Codes and infinite words. Acta Cybernetica 11(4), 241–256 (1994)
4. Devolder, J., Timmerman, E.: Finitary codes for biinfinite words. Informatique Théorique et Applications 26(4), 363–386 (1992)
5. Frey, G., Michel, C.J.: Circular codes in archaeal genomes. Journal of Theoretical Biology 223, 413–431 (2003)
6. Han, N.D., Huy, P.T., Thang, D.Q.: A Quadratic Algorithm for Testing of Omega-Codes. In: Pan, J.-S., Chen, S.-M., Nguyen, N.T. (eds.) ACIIDS 2012, Part I. LNCS, vol. 7196, pp. 338–347. Springer, Heidelberg (2012)
7. Lam, N.H., Van, D.L.: On a class of infinitary codes. Informatique théorique et Applications 24(5), 441–458 (1990)
8. Lam, N.H., Van, D.L.: On strict codes. Acta Cybernetica 10(1-2), 25–34 (1991)
9. Madonia, M., Salemi, S., Sportelli, T.: A generalization of sardinas-patterson algorithm to Z-codes. Theoretical Computer Science 108, 251–270 (1993)
10. Perrin, D., Éric Pin, J.: Infinite Words, automata, semigroups, logic and games. Elsevier Inc. (2004)
11. Van, D.L., Lam, N.H., Huy, P.T.: On codes concerning bi-infinite words. Acta Cybernetica 11(1-2), 97–109 (1993)

Designing an Intelligent Problems Solving System Based on Knowledge about Sample Problems*

Nhon V. Do[1], Hien D. Nguyen[1], and Thanh T. Mai[2]

[1] University of Information Technology, Vietnam
{nhondv,hiennd}@uit.edu.vn
[2] Binh Duong University, Vietnam
thanhmtbdu@gmail.com

Abstract. In computer science and information technology, ontology has been researched and developed in application for knowledge representation. COKB model (Computational Object Knowledge Base) is an ontology was researched and applied in designing knowledge base systems, such as domain of knowledge about analytic geometry, linear algebra. However, when dealing with a practical problem, we often do not immediately find a new solution, but we search related problems which have been solved before and then proposing an appropriate solution for the problem. In this paper, the extension model of ontology COKB has been presented. In this model, Sample Problems, which are related problems, will be used like the experience of human about practical problem, simulate the way of human thinking about finding solution of problem. Besides that, the architect of Intelligent Problem Solving system has been researched and applied. Using this architect and extension COKB model is applied to construct the system for automatic solving problems on knowledge about plane geometry.

Keywords: artifical intelligent, knowledge representation, intelligent system, automated reasoning.

1 Introduction

Knowledge representation play a very important role for designing knowledge base systems as well as intelligent systems. Nowadays there are many various knowledge models which have been suggested and applied. In the books [1 - 4], some knowledge models have been represented in designing knowledge base systems, such as semantic network, neural network, rules of inference and ontologies. Nevertheless, those methods also have several drawbacks which make knowledge representation difficult to implement, especially for areas in general education such as plane geometry, analytic geometry, algebra, physics, etc.

* This research is funded by Vietnam National University HoChiMinh City (VNU-HCM) under grant number C2012-01/HĐ-KHCN.

A. Selamat et al. (Eds.): ACIIDS 2013, Part I, LNAI 7802, pp. 465–475, 2013.

In order to deal with a practical problem, we often do not immediately find a new solution, we search related problems which have been solved before, and then proposing an appropriate solution for the problem. Besides that, we can use the result of related problems for solving the current problem. Such related problems are called *sample problems*. Analogical reasoning which is a method to deduce based on sample problems has been metioned in [12]. However, this method is still general and very difficult to apply in complex real knowledge.

Besides that, applying knowledge about Sample Problems has been researched in Computational Network [10], but this model is only effective for compuational knowledge domain. COKB model in [7], [9] is very useful and suitable for representing knowledge in the domains of reality applications. This model has been used in designing some intelligent systems in education (see [5]). In this paper, we extend COKB model by adding sample problems component to the knowledge of the system.

In [14], the author present structure of Intelligent Problem Solvers (IPS) system and design a process to construct this system. Using extension COKB model and system architecture of IPS, the system for automatic solving problems on knowledge about plane geometry has been built.

2 COKB Model Using Sample Problems

COKB model in [9] is a model which can be presented reality knowledge of human completely and generally. COKB model is more effective than the other knowledge representation methods in many aspects, such as: represtation, inference and communication. With this mdel, we can design the model for general problems and the algorithms which are simulated the the way of human in solving problems based on human knowledge.

2.1 COKB Model and Problem

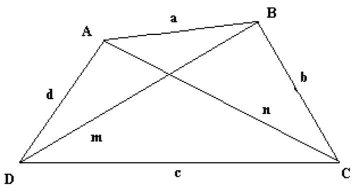
Definition 2.1: The model of computational object knowledge base (COKB model) consists of 6 components:

$$\mathcal{K} = (\mathbf{C}, \mathbf{H}, \mathbf{R}, \mathbf{Ops}, \mathbf{Funcs}, \mathbf{Rules}\ )$$

The meanings of the components are:

- **C** is a set of concepts of computational objects (Com-Object).
- **H** is a set of hierarchy relation on the concepts.
- **R** is a set of relations on the concepts.
- **Ops** is a set of operators.
- **Funcs** is a set of functions.
- **Rules** is a set of rules.

Each concept in C is a class of C-objects. The structure C-Objects can be modeled by (***Attrs, F, Facts, RulesObj***). *Attrs* is a set of attributes, *F* is a set of equations called computation relations, *Facts* is a set of properties or events of objects, and *RulesObj* is a set of deductive rules on facts. For example, knowledge about a Quadangle consists of:



Attrs = {A, B, C, D, a, b, c, d, m, n, S, p, ...} is the set of all attributes of a Quadrangle as following:

A, B, C, D: are objects of ANGLE concept.
a, b, c, d: are sides of quadrangle, they are belong to SEGMENT concept
S, p: are real value about: area, diameter.

F = $\{A + B + C + D = \pi$, p = (a + b + c + d) / 2,
n = a^2+b^2-2.a.b.cos(B), etc.}

Facts = {a+b>n; a+d>m; b+c>m ; etc.}

RulesObj = $\{ \{m \perp n\} \Rightarrow \{S = ½.m.n\}$, etc. }

Definition 2.2: Model of problem on COKB model consists of three sets below.

O = {O1, O2, . . ., On},
F = {f1, f2, . . ., fm},
Goal = { g1, g2, . . ., gm }.

In the above model the set O consists of C-objects, F is the set of facts given on the objects, and Goal consists of goals. A goal of a problem may be the followings:

- Determine an object.
- Determine an attribute (or some attributes) of an object.
- Consider a relation between objects.
- Compute a value of a function relative to objects.

We consider the problem that to determine (or compute) attributes in set **G** from given attributes in set $L = O \bigcup F$. The problem will be denoted by $\mathbf{L \rightarrow G}$

Definition 2.3: Give a problem (O, F, G) or $L \rightarrow G$ on COKB model, M is the set of attributes, $A \subseteq M$.

(a) Each $f \in$ Rules, each $Oi \in O$, we define:

$f(A) = A \bigcup M_A(f)$, with $M_A(f)$ is set of attributes can be deduced from A by f.

$O_i(A) = A \bigcup M_{Oi}(f)$, with $M_{Oi}(f)$ is set of attributes can be deduced from A by O_i

(b) Suppose $D = [r_1,r_2, \ldots, r_m]$ is a list of elemnets which $r_j \in$ Rules or $r_j \in O$. Denote: $L_0 = L$, $L_1 = r_1(L_0)$, $L_2 = r_2(L_1)$, …. $L_m = r_m(L_{m-1})$ and $D(L) = L_m$

A problem $L \rightarrow G$ is called *solvable* if there is a list D such that $G \subseteq D(L)$. In this case, we say that D is a *solution* of the problem.

2.2 Sample Problem

G. Polya has given a way to solve pratical problem in [11]. When dealing with a practical problem, a convenient way to proceed is considering whether we have met a similar or related problem before or not. If so, then the solution for the problem can be obtained effectively. Or we determine whether the result of relating problems can be used to solve the practical problem or not. This leads to a requirement that model of knowledge base needs to be added a new component which can capture this behavior of human. With this adding, the inference engine can find and use results already known quickly.

In paper [15], we have presented a kind of Sample Problem in COKB model. However, this kind has been difficult to simulate the way of human's thinking about practical problem. In this paper, we present the new kind which is necessary additional to model of Sample Problems.

Model of Sample Problem

Definition 2.4: Give knowledge domain $\mathcal{K}$ = (C, H, R, Ops, Funcs, Rules) as definition 2.1, Sample Problem (SP) is a model on knowledge $\mathcal{K}$, it consists of two components as followed:

$$\mathbf{(H_p, D_p)}$$

In which:

H_p is hypothesis of SP. H_p has structure as following:

$H_p = (O_p, F_p)$

O_p: is the set of C-Objects of Sample Problem

F_p: is the set of facts given on objects in O_p

D_p is a list of elements can be applied on O_p. (D_p have been defined in definition 2.3)

Ciriterias of Sample Problem

Base on real domains of knowledge which have been researched, such as knowledge about Geometry (2D-analytical, 3D-analytical and plane geometry), Algebra, Physics (one-way electric, circuit electric), there are two criteria determining if a problem P is Sample Problem:

- First criterion: Frequency of using problem P in the knowledge domain is high. (In the sample space of common problems based on the knowlege domain, frequency of using problem S is greater than some δ)
- Second criterion: Number of Objects in problem P is small (smaller than some λ, it means card(O_p) $\leq \lambda$)

Example 2.1:

Problem P: "Let circle with center O and radius R, AB is a diameter and MB is a chord of this circle."

Hence, we have model of Sample Problem P like this: **P = (Hp, Rp)**

Which $H_p = (O_p, F_p)$

O_p:= { Circle(O, R) , AB: segment, MB: Segment }

F_p:= { AB: DIAMETER of Circle(O), MB: CHORD of Circle(O)}

R_P:={ {AB: diameter of Circle(O)} → {A belongs to Circle(O), O = MIDPOINT(A,B)},

{MB: chord of Circle(O)} → {B belongs to circle (O), M belongs to circle (O)} , {M belongs to circle (O), AB: diameter of Circle(O)} → {ΔMAB is a right triangle} ,

{ΔMAB is a right triangle, O = MIDPOINT(A,B)} → {PO = AB/2}

}

2.3 COKB Model Using Sample Problems

Definition 2.5: Model of Computation objects Knowledge Base with Sample Problems (COKB-SP) consists of 7 components as followed:

(C, H, R, Ops, Funcs, Rules, Sample)

where:

- **(C, H, R, Ops, Funcs, Rules)** is knowledge domain which presented by COKB model.
- **Sample** is a set of Sample Problems of this knowledge domain.

Kinds of facts and model of problem on COKB-SP are defined similarly as definition 2.2.

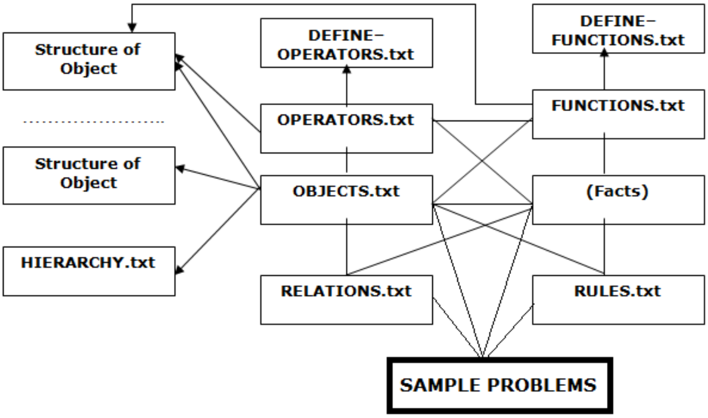


Fig. 1. Structure of COKB-SP model

Model COKB-SP is a useful method for designing practical knowledge bases, modeling complex problems and designing algorithms to solve automatically problems based on a knowledge base. It simulate better for the way of human thinking about searching solution of problem.

3 Algorithms for COKB Model Using Sample Problems

3.1 Algorithm for Finding Sample Problems

Problem S = (O, F, Goal) or S = L → G on model COKB-SP, the Sample Problem P = (H_p, D_p) in Sample set can be found to apply on S by these algorithm:

Algorithm 3.1.1: Problem S and P satisfy two conditions:

(i) Set of Objects of S equal to Set of Objects of P. Denote: $S.O = P.O_p$

(ii) Set of facts of S equal to Set of facts of P. Denote $S.F = P.F_p$

We have algorithm for this case:

```
Select P in Sample
flag←true:
if not($S.O = P.O_p$) then flag ←false:
else if not ($S.F = P.F_p$) then flag ←false:
end if:
if flag then P is SP of S
```

Algorithm 3.1.2: Problem S and P satisfy two conditions:

(i) Set of Objects of S is subset of Set of Objects of P. Denote: $S.O \subset P.O_p$

(ii) Set of facts of S is subset of to Set of facts of P. Denote $S.F \subset P.F_p$

We have algorithm for this case:

```
Select P in Sample
flag←true:
if not($S.O \subset P.O_p$) then flag ←false:
else if not ($S.F \subset P.F_p$) then flag ←false:
end if:
if flag then P is SP of S
```

Algorithm 3.1.3: Problem S and P satisfy two conditions:

(i) Set of Objects of P is subset of Set of Objects of S. Denote: $S.O \supset P.O_p$

(ii) Set of facts of P is subset of equal to Set of facts of S. Denote $S.F \supset P.F_p$

We have algorithm for this case:

```
Select P in Sample
flag←true:
if not($S.O \supset P.O_p$) then flag ←false:
else if not ($S.F \supset P.F_p$) then flag ←false:
end if:
if flag then P is SP of S
```

3.2 Algorithm for Finding Solution of Problem in COKB-SP

Algorithm 3.2: Give the problem S = (O, F, Goal) as definition 2.2 on COKB-SP, the solution of problem S has been found though these steps:

Step 1: Classify problems such as problems as frames, problems of a determination or a proof of a fact, problems of finding objects or facts, etc.

Step 2: Classify facts and representing them.

Step 3: Modeling kinds of problems from classifying in step 1 and 2. From models of each kind, we can construct a general model for problems, which are given to the system for solving them.

The following general algorithm represents one strategy for solving problems: forward chaining reasoning with heuristics and Sample Problems in which objects attend the reasoning process as active agents.

Step 1: Record the elements in hypothesis part and goal part.

Step 2: *Find the* ***Sample Problem*** *can be applied*. (Algorithms 3.1.k , k=1..3)

Step 3: Check goal *G*. If *G* is obtained then goto step 8.

Step 4: Using heuristic rules to select a rule for producing new facts or objects.

Step 5: If selection in step 3 fails then search for any rule which can be used to deduce new facts or new objects.

Step 6: If there is a rule found in step 3 or in step 4 then record the information about the rule, new facts in Solution, and new situation (previous objects and facts together with new facts and new objects), and goto step 2.

Step 7: Else {search for a rule fails} Conclusion: Solution not found, and stop.

Step 8: Reduce the solution found by excluding redundant rules and information in the solution.

4 Application

Structure of Intelligent Problem Solvers (IPS) system and processing to build them has been presented in [14]. Structure of IPS system like this:

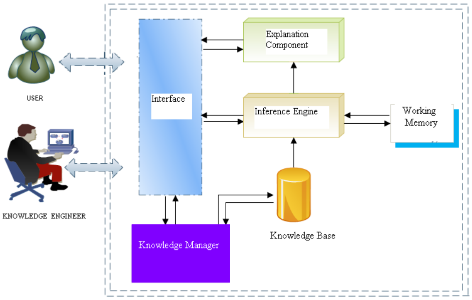


Fig. 2. Structure of IPS system

The main process for IPS: From the user, a problem in a form that the user enter is input into the system, and the problem written by specification language is created; then it is translated so that the system receives the working problem in the form for the inference engine, and this is placed in the working memory. After analyzing the problem, the inference engine generates a possible solution for the problem by doing some automated reasoning strategies such as forward chaining reasoning method, backward chaining reasoning method, reasoning with heuristics. Next, The first solution is analyzed and from this the inference engine produces a good solution for the interface component. Based on the good solution found, the answer solution in human-readable form will be created for output to the user.

Based on structure of IPS system, knowledge base of system has been represented by COKB model using Sample Problems, and Inference Engine has been designed by algorithms 3.1 and 3.2, we built an IPS in education which is called *the system for automatic solving problems in plane geometry*.

4.1 Design the Knowledge Base of Plane Geometry

Base on knowledge about plane geometry in midddle school has been mentioned in [6], this knowledge domain can be represented by model COKB-SP as followed:

a) C–set of concepts of computational objects

The set *C* consists of concepts such as “Point”, “Ray”, “Segment”, “Angle”, “Line”, “Triangle”, “Trapezoid”, “Circle” such as:

- “Point” is a basic object.
- “Ray”, “Segment” are objects of the first level.
- “Triangle” , “Circle is an object of the second level.

b) H–set of hierarchical relation on the concepts

From concepts of the objects introduced above, there are some hierarchical relations on them. Some of them are listed below:

- “Acute Angle”, “Obtuse Angle” or “Straight Angle” is a special “Angle”;
- “Isosceles Triangle”, “Right Triangle ” is also a “Triangle”,

c) R–set of relations on C-Objects

Between C-Objects, there are various kinds of relations. We have some relations:

- <u>Relations of basic level</u>: are relations between basic objects and objects of first level.
- <u>Relations of first level</u>: are relations between basic objects, objects of first level and objects of second level, or relations between objects of higher levels.

d) Ops–set of relations on C-Objects

In knowledge about plane geometry of middle shool, operators are relations between real numbers so we have Ops = { }

e) Funcs–set of functions on C-Objects

The set Funcs consists of functions on C-Objects, such as:

+ Midpoint of a segment.

+ Symmetrical point of a point through a line.

f) Rules–set of rules

Almost properties, clauses, theorems in plane geometry of middle shool can be represented by rules on facts relating to C-Objects. Followings are some particular rules:

{a: segment, b: segment, c: segment, a // b, c ⊥ a} ⟹ {c ⊥ b}

{A: point, B: point, C: point, BC = AC} ⟹ {ABC is an isosceles triangle at C}.

g) Sample–set of Sample Problems:

+ Sample Problems about determining type of Objects, e.g: Right Triangle, Rectangle, Circle

+ Sample problems about:

* Solving Right Triangle.

* Relation between diameter and chord of circle.

* Intersecting chords of a circle.

4.2 Design the Inference Engine

Model of problem in this knowledge base about plane geometry is defined similarly as definition 2.2 as following **(O, F, Goal)**

Besides that, using algorithm 3.1 and 3.2, the inference engine has been built. This engine simulates the way of human thinking about finding solution of practical problem by seraching sample problem can be applied in problem.

Example:

Problem S: Let circle with center O and radius R, AB is a diameter and PB is a chord of this circle. TB is a tagent of circle and the line AP cut TB at T.

Prove: angle PBT = angle APO

+ Specification of Problem:

O:= { Circle(O, R) ;

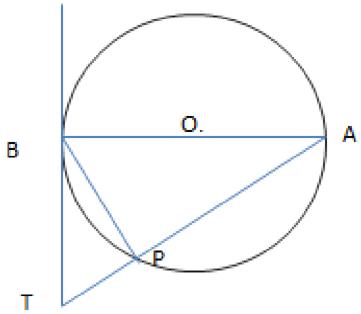
AB, PB, TB: Segment }

F:= { AB: DIAMETER of Circle(O), PB: CHORD of Circle(O),

TB: TAGENT of Circle(O),

P belongs to AT }

Goals:=[“Prove”, Angle(APO)=Angle(PBT)]



+ Solution of Program:

Step 1:

{PB: chord of Circle(O) } → {B belongs to circle (O), P belongs to circle (O)}

By: "Properties of chord “

Step 2:

{AB: diameter of Circle(O)} → {A belongs to Circle(O), O = MIDPOINT(A,B)}

By: "Properties of diameter”

Step 3:

{P belongs to circle (O), AB: diameter of Circle(O)} → {ΔPAB is a right triangle}

By: " Properties of diameter”

Step 4:

{ A belongs to Circle(O), B belongs to Circle(O),
PB: chord of Circle(O), TB: tagent of Circle(O)}
→ {Angle(BAP) = Angle(PBT)}

By: "Properties of tagent and chord"

Step 5:

{ PAB is a right triangle, O = MIDPOINT(A,B)} → {Angle(PAO) = Angle(APO)}

By: "Properties of right triangle"

Step 6:

{Angle(BAP)=Angle(PBT),Angle(PAO)=Angle(APO)}→{Angle(PBT)=Angle(APO)}

Applying properties in step 1 – step 3 is using Sample Problem P in example 2.1 to deduce the new facts of problem S.

5 Conclusion and Future Work

COKB model is very useful and suitable for representing knowledge base, especially knowledge domains about Mathematics, Physics, Chemistry. Moreover, COKB model has been extended by adding a Sample Problem component to the knowledge base and improving deduction techniques on this model, and the extension model is called *COKB using Sample Problem* (COKB-SP). The inference of the system becomes more intelligent and the solution of the system is more natural and similar to human’s.

In the future, we continue to research the COKB-SP model and combine this kind of Sample Problem to the other kind of it in [15] to complete the program for solving problems. Besides that, we apply this model to design IPS system on other knowledge domains, such as Analysis (one variable, multi variables), Physics (mechanics, optical) and Chemistry (organic and inorganic). In addition, the Computational Network and its application has been built successfully in [8]. Thus, the combination of the COKB-SP model and Computational Network is expected to be an effectively model for knowledge representation.

References

1. van Harmelem, F., Vladimir, Bruce: Handbook of Knowledge Representation. Elsevier (2008)
2. Russell, S., Norvig, P.: Artificial Intelligence – A modern approach, 2nd edn. Prentice Hall (2003)
3. Sowa, J.F.: Knowledge Representation: Logical, Philosophical and Computational Foundations. Brooks/Cole (2000)
4. Ertel, W.: Introduction to Artificial Intelligent. Springer (2011) ISSN: 1863-7310
5. Van Do, N.: The architecture of a system for solving problems for learners and design techniques. Scientific Magazine of Education and Technology, Technical teachers' college of Ho Chi Minh City 2(4) (2007)
6. Vietnam Ministry of Education and Training: Textbook and workbook of plane geometry in middle school. Publisher of Education (2010-2011)
7. Van Do, N.: An ontology for knowledge representation and Applications. Proceeding of World Academy of Science, Engineer and Technology 32 (August 2008) ISSN: 2070-70
8. Van Do, N.: Computational Networks for Knowledge Representation, World Academy of Science, Engineering and Technology. In: ICCSISE 2009, Singapore, vol. 56 (August 2009) ISSN 2070 – 3724
9. Van Do, N.: Model for Knowledge Bases of Computational Objects. International Journal of Computer Science Issues (IJCSI) 7(3(8)), 11–20 (2010) ISSN: 1694-0814
10. Van Do, N., Nguyen, H.: Model for Knowledge Representation using Sample Problems and Designing a Program for automatically solving algebraic problems. In: World Academy of Science, Engineering and Technology (ICEEEL 2010), Paris (2010)
11. Polya, G.: How to solve it. Publisher of Education (1997)
12. Sowa, J.F., Majumdar, A.K.: Analogical Reasoning. In: Ganter, B., de Moor, A., Lex, W. (eds.) ICCS 2003. LNCS (LNAI), vol. 2746, pp. 16–36. Springer, Heidelberg (2003)
13. Munakata, T.: Fundamentals of the New Artificial Intelligence: Neural, Evolutionary, Fuzzy and More. Springer-Verlag London Limited (2008)
14. Van Do, N.: Intelligent Problem Solvers in Education: Design Method and Applications. In: Koleshko, V.M. (ed.) Intelligent Systems, pp. 978–953. InTech (2012) ISBN: 978-953-51-0054-6
15. Van Do, N., Nguyen, H.: A Reasoning method on Knowledge Base of Computational Ojects and Designing a System for automatically solving plane geometry problems. In: Do, N., Nguyen, H. (eds.) Proceeding of World Congress on Engineering and Computer Science 2011 (WCECS 2011), San Francisco, USA, pp. 294–299 (October 2011) ISBN: 978-988-18210-9-6

Semantic Representation and Search Techniques for Document Retrieval Systems

VanNhon Do, ThanhThuong T. Huynh, and TruongAn PhamNguyen

Computer Science Faculty, University of Information Technology,
Vietnam National University - Ho Chi Minh City, VietNam
{nhondv,thuonghtt,truonganpn}@uit.edu.vn

Abstract. Nowadays, organizing a repository of documents and resources for learning on a special field as Information Technology, together with search techniques based on domain knowledge or document content is an urgent need in practice of teaching, learning and researching. There have been several works related to methods of organization and search by content. However, the results are still limited and insufficient to meet user's demand for semantic document retrieval. This paper presents a solution for the organization of a repository that supports semantic representation and processing in search. The proposed solution is a model that shows the integration of components such as an ontology describing domain knowledge, a database of document repository, semantic representation for documents and a file system; with problems, semantic processing techniques and advanced search techniques based on measuring semantic similarity. The solution is applied to build a document retrieval system in the field of Information Technology, with semantic search function serving students, teachers, and manager as well. The application has been implemented, tested at the University of Information Technology, Ho Chi Minh City, Vietnam and has achieved good results.

Keywords: document retrieval system, knowledge representation, document representation, semantic search, ontology.

1 Introduction

Electronic libraries and learning resource management systems are indispensable in the application of information technology in education and training. Serving learners, teachers and even managers in satisfying their information need for teaching, learning and researching, these systems are in practical and urgent needs to be increasingly effective but the outcome is still very limited. Popular solutions and technologies have much support for learning resource management, but mainly in the data processing. Various standards for resource description were proposed, e.g. LOM, IMS, Dublin Core, MARC, etc., but these standards are merely used to create metadata with limited specific vocabulary and simple description fields such as title, license, author, etc., not enough to interpret, combine resources by semantic content and thus features of the system is not sufficient to meet the increasing requirements, especially the organization, processing and integration of data, information and knowledge.

A. Selamat et al. (Eds.): ACIIDS 2013, Part I, LNAI 7802, pp. 476–486, 2013.

From the initial, simple Boolean search model, many authors have attempted to improve the efficiency of searching through the more complex models such as Vector Space Model, Probabilistic Models, and Language Model. Many other works that have made effort to change weighting schemes, use natural language processing techniques [8], word sense disambiguation [6], query expansion [1], etc., also contribute to increase search efficiency. Despite many proposals and efforts aimed at improving search results, the limitations of the use of keywords are not overcome yet.

Nowadays, in computer science there is a gradual shift to knowledge orientation or semantic processing. Accordingly, the concept based information retrieval systems have been researched and developed to replace the traditional systems that have revealed several major shortcomings. The search is based upon a space of concepts and semantic relationships between them. Semantic or conceptual approaches attempt to implement some degree of syntactic and semantic analysis; in other words, they try to reproduce, to some degree of understanding of the natural language text that users would provide corresponding to what they think. In particular, the approach based on ontology is considered modern and most appropriate for the representation and handling of content and meaning of documents [2, 4, 5, and 7]. Additionally, richer document representation schemes, proposed by considering not only words but also semantic relations between words such as semantic nets, conceptual graph, star graph, frequency graph, distance graph, etc. are evaluated with high potential: they allow to represent semantic links between concepts whereas poor representation models cannot.

The main goal of this paper is to introduce models, algorithms, and techniques for organizing text document repositories supporting representation, and dealing with semantic information in the search. The paper is organized as follows: section 2 introduces ontology model describing knowledge about a particular field as Information Technology; section 3 presents a graph based document representation model; section 4 introduces a model for organizing, storing document repository on computer; section 5 presents techniques in semantic search; finally, a conclusion ends the paper.

2 Ontology Model

Classed Keyphrase based Ontology model (CK_ONTO) is a system composed of six components:

$$(K, C, R_{KC}, R_{CC}, R_{KK}, \text{label})$$

The components are described as follows:

- Set of keyphrases K

Keyphrase is the main element to form the concept of ontology. In addition, keyphrase also means a structural linguistic unit such as a word or a phrase. There are two kinds of keyphrases: single keyphrase and combined keyphrase. Single keyphrase only represents a concept, formed lexicaly by a single word or a fixed phrase, for example, *computer, network, database, data structure*. Combined keyphrase represents several concepts, formed by a group of single keyphrases that have semantic relationships between components, for example, *computer networking and communication, database programming, network programming*.

Let K = {k | k is a keyphrase of knowledge domain}, $K = K_1 \cup K_2$, in which, K_1 is a set of single keyphrases and K_2 is a set of combined keyphrases.

- Set of classes of keyphrases C

Each class c ∈ C is a set of keyphrases related to each other by certain semantics. A keyphrase may belong to different classes. The classification of K depends on the specialization of concepts. Let C = {c ∈ $\wp$(K) | c is a class of keyphrases which describes the sub topics or sub subjects of knowledge domain}. For example, DATA STRUCTURE class contains keyphrases related to data structures as follows: *DATA STRUCTURE = {stack, queue, contiguous list, linked list, hash table, graph, tree, sorting, strictly binary tree, complete binary tree, AVL tree, Red Black tree, Bubble sort, Merge sort,* etc.*}*

- Set of relations between keyphrase and class R_{KC}

A binary relation between K and C is a subset of $K \times C$ and $R_{KC} = \{r | r \subseteq K \times C\}$. In this paper, R_{KC} only includes a relation called "belongs to" between keyphrase and class, which is defined as a set of pairs (k, c) with k ∈ K, c ∈ C.

- Set of relations between classes R_{CC}

A binary relation on C is a subset of $C \times C$ and $R_{CC} = \{r | r \subseteq C \times C\}$. There are two types of relations between classes are considered: Hierarchical relation and Related relation. A class can include multiple sub classes or be included in other classes. A subclass is a class that inherits some properties from its superclass. The inheritance relationships of classes give rise to a hierarchy or a hierarchical relationship between classes. For instance, *PROGRAMMING LANGUAGE* and *PROGRAMMING TECHNIQUE* are subclasses of *PROGRAMMING.* According to the way to build a class above, a keyphrase may belong to many different classes or a subclass is allowed to have any number of father classes. This leads to the emergence of a relation on which the classes are called "related to each other" but not in meaning of inclusion or containment. These classes have some common properties, more or less related to each other because they have similar keyphrases or subclasses. For example, the related classes are *COMMUNICATION* and *NETWORK*, *HARDWARE* and *ELECTRONIC TECHNOLOGY.*

- Set of relations between keyphrases R_{KK}

A binary relation on K is a subset of $K \times K$, i.e. a set of ordered pairs of keyphrases of K, and $R_{KK} = \{r | r \subseteq K \times K\}$ .There are several different kinds of semantic relations between keyphrases. The amount of relations may vary depending on considering the knowledge domain. These relations can be divided into three groups: equivalence relations, hierarchical relations, non-hierarchical relations.

Equivalence relations link keyphrases that have the same or similar meaning and can be used as alternatives for each other, such as synonym relation, abbreviation relation, near-synonym relation. For example, *JSP* is the short form of *Java Server*

Page, *Twittworking* is synonymous with *Twitter networking*, and *semantic search* is close to *search by content*.

Hierarchical relations, such as "a part of" relation (or part-whole relation), "a kind of" relation (is-a relation), link keyphrases that one of which has a broader (more global) meaning than the other. For example, *soft computing* is a part of *computer science*, *recognition* is a part of *image processing*, *semantic net* is a kind of *graph*, and *Java* is a kind of *programming language*.

Non-hierarchical relations link keyphrases which are semantically related each other without forming a hierarchy or semantic equivalence, such as Expansion, Same-class, Cause, Influence, Instrument, Make, Possession, Source, Aim, Location, Temporal, Manner, Support, Beneficiary, Property, Agent, Circumstance and Person.

- Labeling function for classifying keyphrase

A keyphrase may refer to a terminology or a class to which the keyphrase belongs and its name is the same as name of the class. Thus, the semantics of a keyphrase may relate to its level of content (or level of its class) such as discipline, major, subject, theme, topic. To describe the information that a keyphrase represents a class and level of the class, a labeling function is used. For example, *soft computing* ↦{"terminology", "major"}.

3 Document Representation

Understanding the document content involves not only the determination of the main keyphrases occur in that document but also the determination of semantic relations between these keyphrases. Therefore, each document can be represented by a graph of keyphrases in which semantic relations connect keyphrases to each other.

Definition: A Keyphrase Graph (KG) defined over ontology CK_ONTO, is a triple (G_K, E, l) where:

- $G_K \subset K$ is the non-empty, finite set of keyphrases, called set of vertices of the graph.
- E is a finite set with elements in $G_K \times G_K$, called set of arcs of the graph. The arc is always directed and represents a semantic relation between its two adjacent vertices.
- l: $E \rightarrow R_{KK}$ is a labeling function for arcs. Every arc $e \in E$ is labeled by relation name or relation symbol.

When these graphs are used for representing a document, keyphrase vertices represent keyphrases of CK_ONTO ontology treated in the document (reflect the main content or subject of the document), and the labeled arcs represent semantic links between these keyphrases. For example:

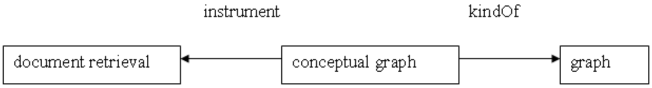


Fig. 1. An example of keyphrase graph

Definition: An extensive keyphrase graph, denoted as G_e, derived from keyphrase graph G = (G_K, E, l), is a triple (G_K, G_R, E') satisfying the following conditions:

- (G_K, G_R, E') is a bipartite, finite and directed graph.
- $G_K \subset K$ is a non-empty keyphrase vertex set.
- $G_R \subset R_{KK}$ is a relation vertex set which represents the semantic relations between keyphrases. (The vertex set of the graph is N =$G_K \cup G_R$, $G_K \cap G_R \neq \emptyset$). Each arc $e \in E$ is correspond to a vertex $\tilde{r} \in G_R$ with $\tilde{r}$ = (e, lab(e)).
- E' is a non-empty set with elements in $G_K \times G_R \cup G_R \times G_K$, called set of arcs of the graph. Vertices of the bipartite graph are divided into two nonempty, disjoint sets G_K and G_R, with two different kinds of vertices. All arcs then connect exactly one vertex from G_K and one vertex from G_R. Therefore, all arcs go from a keyphrase vertex either to a relation vertex or from a relation vertex to a keyphrase vertex.

This extensive keyphrase graph can be considered a variant of conceptual graph. There is 1:1 correspondence between a keyphrase graph and its extensive form. We can be easily transformed from the original keyphrase graph to the extensive graph and vice versa. Using which form of graph depends on its convenience in representation, storage, processing, calculation or implement. The illustration below is the same keyphrase graph in Figure 1 but in extensive form:

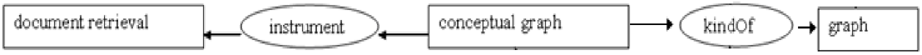


Fig. 2. An extensive keyphrase graph

Definition: Let G = (K, R, E) be a keyphrase graph (in extensive form). A sub keyphrase graph (subKG) of G is a keyphrase graph G' = (K', R', E') such that: $K' \subseteq K, R' \subseteq R, E' \subseteq E \; and \; (i,j) \in E' \Rightarrow i,j \in K' \cup R'$.

A subKG of G can be obtained from G only by repeatedly deleting a relation vertex (and arcs incident to this vertex) or an isolated keyphrase vertex.

4 Semantic Document Base Model

This section considers a model of organizing a document repository on computer that supports tasks such as accessing, processing and searching related to document content or the semantics. This model is called "Semantic Document Base" model (SDB model). A SDB model is a system composed of five components, denoted by:

(D, FS, DB, CK_ONTO, SBD_R),

in which the components are described as follows:

- Set of documents D

This is a collection of real document not classified or handled. Each document $d \in D$ has physical representation in the storage system as a file. However, in practice there are many documents stored in some files, i.e. that each document can include several sections each of which is stored into a separate file, but in scope of this research, each document is considered as a file.

- Model of file system of document repository FS

Storage system is organized as a hierarchical system of directories, or hierarchical directory tree. A directory can contain many subdirectories or documents, whereas each directory or document can have one parent directory only. A distinctive feature of the FS is that naming directories and organizing the directory hierarchy and classifying documents into directories must follow some predefined rules described as follows:

Directory Naming rule: directory name must be normalized by the keyphrase representing a certain class defined in the CK_ONTO. That is, each directory corresponds to a class in the ontology describing the sub topic in the knowledge domain.

Hierarchical Organization rule: The hierarchy between directories must follow the hierarchy of classes in the CK_ONTO. For the directory system of learning documents, the hierarchy is made from the wide range such as discipline and major to the narrower range such as courses, subjects or topics.

Rule of classifying documents into directories: Each document is represented by a list of keyphrases that describe major topics of the document, and each directory is also named by a keyphrase expressing semantic information. Measuring the semantic similarity between keyphrases representing directories and keyphrases representing a document gives a way of classifying the document into a corresponding directory.

- Model of database of document repository DB

The database of document repository is created based on the relational database model and Dublin Core standards. Besides the common elements of Dublin Core, each document includes some own special features and attributes to express its information structure in more detail and semantic information related to document content. For example, the information structure of the thesis includes own features such as scientific advisors, thesis defense committee and marks.

- An ontology describing domain knowledge CK_ONTO

The ontology model describes knowledge of the domain (as presented in Section 2) is a knowledge representation model for a special domain, including six components: (1) the set K of keyphrases, (2) the set C of classes of keyphrases describing sub subjects in the knowledge domain, (3) the set R_{KC} of relations between the keyphrase and class, (4) the set R_{CC} of relations between classes, (5) the set R_{KK} of relations between the keyphrases, and finally a labeling function used for classifying keyphrase based on its level of content.

- Set of relationships between components SDB_R

All relationships between the components in the SDB model called Semantic Document Base _ Relationship (SDB_R) includes:

1/. Each document d ∈ D is stored in a unique directory of the FS system, which determines a mapping:
$pos: D \rightarrow FS$
$d \mapsto pos(d)$, for each document d ∈ D, there is a path pos(d) referring to a node on the FS directory tree.

2 /. Each document d ∈ D has a record in the database DB.
$record: D \rightarrow r(DOCUMENT) \in DB$
$d \mapsto record(d) = t$

Each tuple t of the relation r (DOCUMENT) stores information of a real document d with title, author, keywords, description, the name of the physical file…and the attribute idDocument is used as the primary key to distinguish one document from another.

3/. Each document d ∈ D is represented by a keyphrase graph KD(d) ∈ F_{KG} (F_{KG} is a set of keyphrase graphs) in which keyphrase vertices represent keyphrases of CK_ONTO treated in the document and relation vertices represent semantic relations between these keyphrases.
$KG: D \rightarrow F_{KG}$
$d \mapsto KG(d)$

4/. Each directory in FS corresponds to a class in ontology CK_ONTO and the hierarchical relation between directories depends on the hierarchical relation between classes of the ontology. Then, there is a mapping:
$cl: X \rightarrow C$
$x \mapsto cl(x)$ so that for all x, y ∈ X, if x f y then cl (y) ⊂ cl (x),
in which X is the set of directory names and f is the hierarchical relation.

5/. Each keyphrase k ∈ K is assigned a weight as idf (Inverse Document Frequency)
$idf: K \rightarrow [0, \log(|D|)]$

6/. For the document d ∈ D, each specific keyphrase is assigned two local weights such as tf (Term Frequency) and ip (Importance of Position), attempting to measure the importance of the keyphrase within a given document
$d \rightarrow (KG, tf: V(KG) \rightarrow [0,1], ip: V(KG) \rightarrow [0,1])$, where V(KG) is the vertex set of KG(d).

5 Semantic Search

This section will discuss an approach for semantic search based on relevance evaluation between the target query and documents by calculating measures of semantic similarity between keyphrases, relations and keyphrase graphs representing documents. The definitions of semantic similarities are given based on ideas of D. Gennest and M. Chein [3] with some modifications.

5.1 Weighting for Keyphrases

To improve the representational power of keyphrases as a descriptor of the given document, weights are assigned to them. Meanwhile, each document will be represented by a weighted keyphrase graph more semantically rich. One of the most successful and commonly used techniques in information retrieval is to use Term Frequency - tf and Inverse Document Frequency - idf.

The “Term frequency” (tf) is the frequency of occurrence of the given keyphrase within the given document. Frequency of occurrence of keyphrase k in document d, denoted as tf(k,d) is defined as follows:

$$tf(k,d) = \frac{n_k}{\sum_{i \in d} n_i} \tag{1}$$

where n_k is the number of occurrences of the keyphrase k in the document d.

The “Inverse document frequency” (idf) characterizes a given keyphrase within an entire collection of documents. It is a measure of how widely the keyphrase is distributed over the given collection, and hence of how likely the keyphrase is to occur within any given document by chance. Value of idf(k) represents the specificity of keyphrase k in collection D and computed by:

$$idf(k) = log\left(\frac{|D|}{1 + |\{d \in D, k \in d\}|}\right) \tag{2}$$

Document information structure composed of many separate components and each component has different meaning, role, and location in the semantic descriptions or document content. Therefore, each keyphrase has different importance when they appear in the different components of the document content. For example, the title shows brief information in the document, acts as a gateway to the content that helps readers quickly grasp the general meaning of the text. So the keyphrases appear in the title is always considered to be significant and have the highest priority. The weight called ip (Importance of Position) is considered to represent an estimate of the importance or usefulness of the given keyphrase according to the location where it appears in document. The importance of keyphrase k in document d based on the location of occurrence of k in the document, denoted as ip(k,d) is defined as follows:

$$ip(k,d) = \frac{\sum_i w_i n_i}{\sum_i n_i} \tag{3}$$

where, w_i is the weight representing an estimate of the importance of i^{th} component of document structure and n_i is the number of occurrences of the keyphrase k in this component, with constraints as $w_i \in [0,1]$ and $\sum_i w_i = 1$.

5.2 Relevance Evaluation

A keyphrase graph is constituted by keyphrases and relations, so the direction to measure semantic similarity between graphs is to calculate the similarity between keyphrases and the similarity between relations used in the graphs.

Let $\alpha: K \times K \to [0,1]$ and $\beta: R_{KK} \times R_{KK} \to [0,1]$ be two mappings to measure semantic similarity between two keyphrases and two relations defined in the

CK_ONTO ontology. 1 represents the equivalence between two objects and 0 corresponds to the lack of any semantic link between them. The values of β are selected manually based on the opinions of experts of the field. Determining manually the values of β is possible because of the small number of relations.

Definition: Let k, k' ∈ K, a binary relation P on K defined as : P (k,k') iff k = k' or $\exists S = (s_1, s_2, \dots, s_n)$ a sequence of integers ∈ [1, t] (t = |R$_{KK}$|) such that $k\, r_{s_1} k_1, k_1\, r_{s_2} k_2, \dots, k_{n-1}\, r_{s_n} k'$ with r$_i$ is a relation of R$_{KK}$ (for all x and y in K, x has a relation r with y if and only if (x, y) ∈ r, written as x r y).

The mapping α may be defined by using the sequence used in the relation P as follows:

$\alpha(k, k') = 0\ if\ notP(k, k')$
$\alpha(k, k') = Max\{V(k\, r_{s_1} k_1, k_1\, r_{s_2} k_2, \dots, k_{n-1}\, r_{s_n} k')\}\ if\ \exists S = (s_1, s_2, \dots, s_n)$ a sequence of integers ∈ [1, t] (t = |R$_{KK}$|) such that $k\, r_{s_1} k_1, k_1\, r_{s_2} k_2, \dots, k_{n-1}\, r_{s_n} k'$. (4)

Mapping V allows considering the various semantic relations used in the sequence, is defined as follows:

$$V\left(k\, r_{s_1} k_1, k_1\, r_{s_2} k_2, \dots, k_{n-1}\, r_{s_n} k'\right) = xidf(\mathrm{k}) \prod_{1}^{n} val_{r_{s_i}}(k_{i-1}, k_i) xidf(k_i) \quad (k_n \equiv k')(5)$$

, in which, xidf(k) is a function that has the same meaning as idf (k), but returns a value in the range [0,1], $xidf(k) = idf(k)/\log(|D|)$ and $\mathrm{val}_{r_{s_i}}(k_{i-1}, k_i)$ is the weight assigned to relations r_{s_i} over pair of keyphrases (k$_{i-1}$, k$_i$). This weight is a measure of semantic similarity between the keyphrases k$_{i-1}$ and k$_i$ that are directly linked by the relationr_{s_i}. The value of $\mathrm{val}_{r_{s_i}}(k_{i-1}, k_i)$ is determined using expert method.

The mapping V allows to evaluate the combination of semantic relations used in sequence. This is necessary because the semantic similarity between two keyphrases linked by a semantic relation may vary depending on the used relation. Some links represent a large difference in meaning while other links represent small semantic distance. For example, keyphrases linked by a synonym relation are more semantically likeness than keyphrases linked by a hierarchical relation. Moreover, pairs of keyphrases linked by the same relation may have different semantic similarity. For instance, in a hierarchy tree, the links closer to the root node often have greater semantic distance than the lower-level links. If there may exists many sequences from k to k', value of α(k, k') depends on the maximum of V.

Definition: Let H = (KH, RH, EH) and G = (KG, RG, EG) be two keyphrase graphs defined over CK_ONTO. A projection from H to G is an ordered pair $\prod = (f, g)$ of two mappings $f: RH \to RG, g: KH \to KG$ satisfying the following conditions:

- Projection preserves the relationships between vertices and arcs, i.e. for all $r \in RH, g\big(adj_i(r)\big) = adj_i(f(r))$, adj$_i$(r) denotes the ith vertex adjacent to relation vertex r.
- $r \in RH, \beta\big(r, f(r)\big) \neq 0$
- k $\in KH, \alpha(k, g(k)) \neq 0$

Below described formula allows valuation of one projection. The purpose of a searching method is the semantic resemblance calculus between a query and an document, therefore in valuation formula of the projection from H to G, H is a query graph and G is a document graph.

Definition: A valuation pattern of a projection $\prod = (f, g)$ from a keyphrase graph H to a keyphrase graph G is defined as follows:

$$v(\Pi) = \frac{\sum_{k\in \mathrm{KH}} tf(g(k), G) * \alpha\big(k, g(k)\big) * ip(g(k), G) + \sum_{r\in RH} \beta\big(r, f(r)\big)}{(|KH| + |RH|)} \qquad (6)$$

Figure 3 shows the indexation of the Document #30 and the projection from the Query 1 with relevance ratio 46%

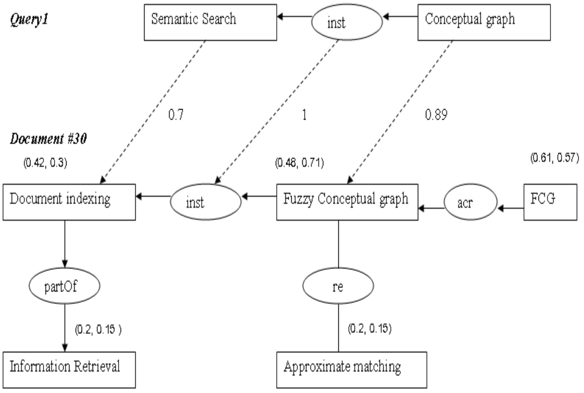


Fig. 3. Matching keyphrase graph

There is a partial projection from a keyphrase graph H to a keyphrase graph G iff there exists a projection from H', a sub keyphrase graph (subKG) of H, to G. A valuation pattern of partial projection $v(\Pi_{partial})$ only depends on vertices of H' and is defined like projection $v(\Pi)$.

Based on valuation of projections, semantic similarity between two keyphrase graphs calculus is defined as follows:

$$Rel(H,G) = Max\{v(\Pi) | \Pi \ is \ a \ partial \ projection \ from \ H \ to \ G\} \tag{7}$$

Determining if a document is relevant for a user query and estimate this relevance is done by calculating the semantic similarity between the keyphrase graphs that represent them.

Conclusion

A solution for the organization of a semantic document repository that supports semantic representation and processing in search is described. The proposed solution is a model that shows the integration of components such as an ontology of the relevant domain, a database, semantic representation for documents and a file system; with semantic processing and searching techniques. The solution is applied to build a document retrieval system with semantic search function. The application has been implemented, tested at the University of Information Technology Ho Chi Minh City, Vietnam and search results have been highly appreciated by users. Recall and precision are used to evaluate the effectiveness of the document retrieval systems. Document repository for testing consists of 10,000 documents as papers, books, slides, essays, etc., evenly distributed into five branches of Information Technology such as Computer Science, Information Systems, Networking and Communication, Computer Engineering, Software Engineering. According to test results on 200 selected queries, the average precision of the system is 87.16% and the average recall is 88, 32%. The research results will be the basis and tools for building many resource management systems in various different fields.

References

1. Aly, A.A.: Using a query expansion technique to improve document retrieval. International Journal Information Technologies and Knowledge (2008)
2. Bonino, D., Corno, F., Farinetti, L., Bosca, A.: Ontology Driven Semantic Search. WSEAS Transaction on Information Science and Application 1(6), 1597–1605 (2004)
3. Genest, D., Chein, M.: An Experiment in Document Retrieval Using Conceptual Graph. In: Lukose, D., Delugach, H., Keeler, M., Searle, L., Sowa, J. (eds.) ICCS 1997. LNCS, vol. 1257, pp. 489–504. Springer, Heidelberg (1997)
4. Styltsvig, H.B.: Ontology-based Information Retrieval. A dissertation Presented to the Faculties of Roskilde University in Partial Fulfillment of the Requirement for the Degree of Doctor of Philosophy (2006)
5. Eriksso, H.: The semantic-document approach to combining documents and ontologies. International Journal of Human-Computer Studies 65(7), 624–639 (2007)
6. Stokoe, C., Oakes, M.P., Tait, J.: Word sense disambiguation in information retrieval revisited. In: Annual ACM Conference on Research and Development in Information Retrieval Toronto, Canada (2003)
7. Tran, T., Cimiano, P., Rudolph, S., Studer, R.: Ontology-Based Interpretation of Keywords for Semantic Search. In: Aberer, K., Choi, K.-S., Noy, N., Allemang, D., Lee, K.-I., Nixon, L.J.B., Golbeck, J., Mika, P., Maynard, D., Mizoguchi, R., Schreiber, G., Cudré-Mauroux, P. (eds.) ISWC/ASWC 2007. LNCS, vol. 4825, pp. 523–536. Springer, Heidelberg (2007)
8. Vallez, M., Pedraza-Jimenez, R.: Natural Language Processing in Textual Information Retrieval and Related Topics. I.S.S.o.t.P.F. University (2007)

Opposition Differential Evolution Based Method for Text Summarization

Albaraa Abuobieda[1,2], Naomie Salim[1],
Yogan Jaya Kumar[1], and Ahmed Hamza Osman[1,2]

[1] Faculty of Computer Science and Information System,
Universiti Teknologi Malaysia, 81310, Skudai, Johor, Malaysia
[2] Faculty of Computer Studies,
International University of Africa, 2469, Khartoum, Sudan
{albarraa,ahmedagraa}@hotmail.com, naomie@utm.my, yogan@utem.edu.my

Abstract. The Evolutionary Algorithms (EAs) save sufficient data about problem features, search space, and population information during the runtime. Accordingly, the machine learning (ML) techniques were employed for examining these data to improve the EAs search performance compared with their classical versions. This paper employs an Opposition-Based Learning as ML approach for enhancing the initial population of the Differential Evolution algorithm in problem of text summarization. In addition, it investigates the use of the OBL technique in integer-based evolutionary populations. The objective of this proposed enhancement is to adjust the algorithm booting instead of relying on random numbers generations only. Basically, all methodology steps in this paper were presented by a previous study whereas the differences between both of them will be shown later. So, this paper tries to estimate the improvement size the OBL can achieve and compare the results with a traditional DE-based text summarization application and other baseline methods. The DUC2002 data set was assigned as a test bed and the ROUGE toolkit used to evaluate the methods performances. The experimental results showed that our proposed method assured the need for learning and improve the random-based EAs before proceed generating the solutions. The study findings conclude that our proposed method outperformed a classical DE and other baseline methods in terms of F-measure. OBL was broadly tested before in numerical test beds, in this paper it will be tested on text-based test bed news article of text summarization problem.

Keywords: Differential Evolution, Evolutionary Algorithms, Text Summarization, Opposition Base Learning, OBL, Machine Learning, DUC, Diversity, Cluster.

1 Introduction

The automatic text summarization (ATS) is an application generates a condensed form of input article in a relevant manner. Achieving such extractive

A. Selamat et al. (Eds.): ACIIDS 2013, Part I, LNAI 7802, pp. 487–496, 2013.

summaries is not so much easy task; the designed system should carefully select sentences that lead to a diverse summary. Some of the proposed methods were built using evolutionary algorithms (EAs) to act as a machine learner for problems such as feature weighting, classification and clustering. There are several text summarization studies designed using EAs. To the best of our knowledge, the literature showed that none of the optimized based ATS methods employed machine learning approach for enhancing the search performance of the EAs themselves before utilizing them. So the contributions of this work are represented in; 1) the OBL was never presented separately or joined with EAs to solve a problem of ATS; 2) all OBL previous studies were not investigated to learn integer-based populations; 3) all OBL previous studies, separately or joined with EAs, were only trained and tested in numerical test bed but not in text-based test bed. Mainly, in this study we implemented a work proposed by [1]. The differences between this paper and the latter are as follows. In addition to the three main contributions declared above, we used the feature based approach for extracting top scored summary sentences instead of sentence-centroid approach which used by the previous work. There are several numbers of machine learning (ML) techniques used for enhancing the search performance of the evolutionary computing (EC) algorithms and abbreviated to (MLEC). The enhancement is can be done on the different stages such as the initial population generation, crossover, mutation, individual selection and so on. For more information regard improved EAs studies using the MLEC, the reader can refer to the recent survey which simply covers these issues [2]. In the literature, we found that most of the ATS applications that were built using the evolutionary algorithms are relying on random number base generation; the selection of text summarization studies using EAs such as Genetic algorithm, particle swarm optimization and Differential Evolution can be found in [3,4,5], respectively. The literature showed that EAs, and in particular the DE, were widely optimized using Opposition-Based Learning (OBL) and achieved better solution than their classical versions [6].The opposition-based computation was mathematically proved better and closer to the optimal goal than the random solution [6]. Therefore, in this paper the OBL proposed to be used in order to optimize a DE-based ATS application. Practically, the ML approach can be applied over all evolutionary algorithm stages. In this paper, only the initial population of the DE algorithm [7] will be enhanced using the OBL machine learning approach. Initial population enhancement includes two sub-stages: Opposition based Population Initialization and Opposition-based Generation Jumping. Section 2 gives more details on OBL theory and mathematical concept. The reason that led us to conduct this experiment is as follows. After we implemented the proposed method [1] we noted that when we rerun the method for several times, at each run we got a different result. Some of them are close to the model summaries and the others are too far from the model summaries. The different outputs for the same inputs are resulted in generate different clusters; thus we induced that when the DE boots using random number generation that was led it to generate disqualified clusters. Critically, the initial population is responsible for allocating the initial sentence clustering distributions. So, we thought to rebuild the application through adjusting the boot population using the OBL approach. The challenges addressed in this paper are as follows. First, can the OBL technique being used to optimize information retrieval-based applications such as summarization? Second, what is the likelihood success of implementing the OBL in an integer-base population? Third, is the OBL technique able to optimize a text-based test bed problem? The OBL was widely successfully implemented in numerical-based test bed [6]. This paper is organized as follows. Section 2 discusses the whole proposed methodology. Then, the experimental findings will be presented in Section 3. Lastly, Section 4 concludes the paper.

2 The Methodology

Subsections 2.1 till 2.6 describe our basis methodology. The four main subsections are the OBL, the DE configuration steps, the similarity measure, the selected features and the used dataset and the evaluation toolkit. To test the validity of the clustered sentences, a summary is generated and evaluated using ROUGE toolkit [8].

2.1 The Opposition-Based Learning

Principally, all EAs family [2] set initialize their populations based on random real values. For example, DE initializes its population in a random manner as in Eq. 1.

$$X_{(i,j)} = a_j + rand_j(0,1) \times (a_j - b_j) \tag{1}$$

Where $X_{(i,j)}$ is the candidate solution vector in generation i, $i = 1, 2, \ldots,$ N_p, and N_p is a population size; $j = 1, 2, \ldots, D$ (D: problem dimensionality); a_i and b_j are upper and lower interval boundaries of the j^{th} variable, respectively, and rand(0, 1) is a uniformly random number in [0, 1]. Enhancing the initial population is a process of presenting individuals that are considered much closer to the optimal solutions than their classic version (random guesses). Then, both individual types are evaluated for best selection. The fittest chromosome is will be selected and survive for the next round. Equation 2 is used to generate the opposite candidate solutions of the vector $X_{(i,j)}$:

$$\widehat{X}_{(i,j)} = a_j + b_j - X_{(i,j)} \tag{2}$$

where *widehatX* is the opposite point of the previous generated point X as in Eq. (1). Next, a non updating of the interval search [a, b] at the next populations would result in searching outside the reduced search space [9]. That is due to the missed knowledge about the updated search space (converged population). For this reason, the interval variables [a,b] would be dynamically updated to $[Min_j^P, Max_j^P]$ where Min and Max are minimum and maximum value of variable j at current population P. Mainly, the OBL technique was proposed to learn the EAs through their initial population and the upcoming generations. The following sub points discuss these issues.

Opposition-Based Population Initialization

To generate an initial population using the OBL technique:

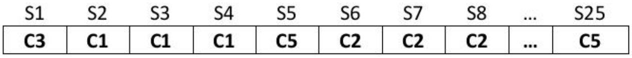
- Randomly initializes a population $P(N_p)$, see Eq. 1.
- Generates Opposite of initial Population $OP(N_p)$, see Eq. 2.
- Calculates fitness functions of $f(P(N_p))$ and $f(OP(N_p))$, see section 2.4.
- Selecting N_p fittest individuals from $\{P(N_p) \cup OP(N_p)\}$ and then establish the initial population.

Opposition-Based Generation Jumping

To produce the new generation of population $P(N_p)$, this step is similar to the previous one except dealing with updating the new interval variables. This is not to allow the algorithm searching out of the new dimension. The jumping rate factor J_r is used as a threshold to generate opposite-based chromosome; the previous studies proved that the best value of the J_r factor is 0.3. So, only a corresponding random variable with less than this threshold its opposite number will be generated, and measured for survival compete. One of the main challenges of this study is how to deal with the mutation and population of an integer-coded based as presented in [1]; the OBL technique has only presented for real-coded based problem [6]. In this paper we decided not to activate the J_r factor for two reasons. First, there weren't numerical related works dealt with integer-base jumping factor before. Hence, our method is continuing jumping into the next generation without using the J_r variable. Second, updating such this variable J_r may not considered an optimal choice for many kind of applications. For example, this method deals with integer discrete values used to assign each sentence to a cluster. So, modifying the interval in this case affects the number of required clusters needed, where the interval variables represent the minimum and maximum cluster labels required.

2.2 The Chromosome Representation

A chromosome representation concerns on how to represent and formulate a specific problem. In this study, the chromosome represents a full document and the number of genes in the chromosome represents the number of sentences of that document in the same order of the genes position. A gene takes a value between $[1, k]$ where k is the maximum number of required clusters. Figure 1 visualizes the chromosome representation and encoding. ; where S_i refers to the



S1	S2	S3	S4	S5	S6	S7	S8	...	S25
C3	C1	C1	C1	C5	C2	C2	C2	...	C5

Fig. 1. A sample of a chromosome representation

sentence $S \in (1, \ldots, n)$ and n is total number of the sentences in the document; C_j refers to the cluster label where $C \in (1, .., j)$ and j refers to total number of required clusters. This chromosome is a representation of a document consists of 25 sentences. The figure tells us to place sentences 2, 3 and 4 into the cluster number 1.

2.3 The Mutation Operation

In this subsection, we discuss the mutation operation that was proposed by [1]. Unlike the traditional DE, this proposed DE method initializes its population using random integer values. The interval of these integer values (genes) ranges between $[1, k]$. An example of an integer chain of the population is $X_r(t) = [X_{(r,1)}(t), X_{(r,2)}(t), ..., X_{(r,n)}(t)]$ where the chromosome $X_{(r,s)}(t) \in \{1, 2, ..., k\}$, $r = 1, 2, ..., N$ and $s = 1, 2, \ldots, n$; N is the population size, and n is the number of sentences in the document. Consider the following example. A document $D = [S_1, S_2, ..., S_{20}]$ consists of 20 sentences of length $n = 20$, a number of required clusters $k = 4$, and the population size $NP = 4$. Then a population can be represented as follows: $X_1(t) = [3, 4, \ldots, 1]$, $X_2(t) = [3, 1, \ldots, 2]$, $X_3(t) = [1, 4, ..., 2]$, and $X_4(t) = [2, 2, \ldots, 4]$. From the previous representation and the given inputs, chromosome $X_1(t)$ commanded the following: allocate sentences S_1 into cluster C_3,S_2 into C_4, and S_{20} into C_1. The same goes with other chromosomes $X_2(t)$, $X_3(t)$, and $X_4(t)$. To generate the mutant vectors $Y_r(t+1)$, the DE needs to randomly select three individuals $X_{r1}(t)$, $X_{r2}(t)$, and $X_{r3}(t)$ where $r1 \neq r2 \neq r3$. Then computes the difference between $X_{r2}(t)$ and $X_{r3}(t)$, scales it using the control parameter F, sums the scaled result to $X_{r1}(t)$. Finally, composes the trail offspring $Y_r(t+1) = [Y_{(r,1)}(t+1), Y_{(r,2)}(t+1), \ldots, Y_{(r,n)}(t+1)]$. In the previous work[1], the DE control parameters assignment were not explained, so Section 2.4 shows how did we assign them. To enable the DE dealing with this explained integer format as in the example above, a genetic mutation operator was adopted according to formula 3:

$$Z_{(r,s)}(t+1) = \begin{cases} 1 & \text{if } rand(s) < sigm(Y_{(r,s)}(t+1)) \\ 0 & otherwise \end{cases} \tag{3}$$

The adopted vector $Z_s(t+1)$ holds the required changes to move the particle $X_r(t)$ to the new position $X_r(t+1)$. Now there are two probabilities: either a gene on vector $Z_{(r,s)}(t+1)$ holds (0) or (1). If it includes 1, then copy and move gene $X_r(t)$ to the new position $X_r(t+1)$. Otherwise, the gene will be mutated. The next example shows in steps how a mutation can be occurred. Consider the following mutation steps, vector $X_r(t)$ needed to move to the new position $X_r(t+1)$ where:

1. $X_r(t) = [3, 2, 4, 23, 1, 4, 1]$
2. $Y_r(t+1)$:
   (a) $Y_{(r,s)(t+1)} = \begin{cases} X_{(r1,s)}(t) + F\left(X_{(r2,s)}(t) - X_{(r3,s)}(t)\right), & \text{if } rand(s) < CR \\ X_{(r,s)}(t), & otherwise \end{cases}$

(b) $Z_{(r,s)}(t+1) = \begin{cases} 1, & \text{if } rand(s) < sigm\left(Y_{(r,s)}(t+1)\right) \\ 0, & otherwise \end{cases}$

3. $X_r(t+1)$:
   (a) If $Z_{(r,s)}(t+1) = 1$, then $X_{(r,s)}(t+1) = X_{(r,s)}(t)$
   (b) If $Z_{(r,s)}(t+1) = 0$, then $Z_{(r,s)}(t+1) = 0 = \{1,4,6,7\}$ // genes addresses not genes values
   (c) $S^{+} = \max S = \max\{1,4,6,7\} = 7$
   (d) $S^{-} = \min S = \min\{1,4,6,7\} = 1$
4. $X_{(r,s^{-})}(t+1) = X_{(r,s^{+})}(t) : X_{(r,1)}(t+1) = X_{(r,7)}(t) = 4$
5. $X_(r,s^{+})(t+1) = X_(r,s^{-})(t) : X_(r,7)(t+1) = X_(r,1)(t) = 3$
6. $S = \frac{S}{\{S^{+},S^{-}\}} = \frac{\{1,4,6,7\}}{\{1,7\}} = 4,6$
7. Then, go to step (2), until $S = \frac{S}{\{S^{+},S^{-}\}} = \{\varnothing\}$

2.4 Objective Function

In order to adjust the quality of the partitional clustering, the same objective function that had been proposed in [1] was used in this paper. The adapted objective function, which is a combination of two criterion functions, is used to balance both intra-cluster similarity and inter-cluster dissimilarity. Intra-Cluster Similarity: This criterion function is used to adjust the similarity degree between the grouped sentences in the given cluster. The maximum similarity score obtained, the one is much higher required. Eq. 4 shows how to compute the intra-cluster similarity.

$$F1 = \sum_{l=1}^{k} |C_l| \sum_{S_i,S_j \in C_l} sim_{NGD}(S_i,S_j) \to \max \tag{4}$$

Where C is a cluster, l is the cluster number, k is the total number of clusters, sim is a similarity, NGD is the current similarity measure selected. S_i and S_j are two sentences currently selected to measure the similarity degree between them and to report the consistency of the cluster itself.

Inter-Cluster Dissimilarity: This criterion function is used to adjust the dissimilarity degree between the cluster sentences in the given cluster. The minimum similarity score obtained, the one is much higher required. Eq. 5 shows how to compute the inter-cluster dissimilarity.

$$F2 = \sum_{l=1}^{k-1} \frac{1}{|C_l|} \sum_{m=l+1}^{k} \frac{1}{|C_m|} \sum_{S_i \in C_l} \sum_{S_j \in C_m} sim_{NGD}(S_i,S_j) \to \min \tag{5}$$

Where C is a cluster, l and m are the clusters, k is the total number of clusters, sim is similarity, NGD is the current similarity measure selected. S_i and S_j are two sentences currently selected to measure the dissimilarity degree between the two clusters. The following objective function, as shown in Eq. 6, is designed to draw the balance of both inter-cluster dissimilarity and intra-cluster similarity.

$$F = \left(1 + sigm(F_1)\right)^{F_2} \to max \tag{6}$$

Where $sigm(n)$ is a sigmoid function used to modulate the objective function value in a binary range $[0, 1]$. The $sigm$ function can be computed as in Eq. 7.

$$sigm(n) = \frac{1}{1 + exp(-F_1)} \tag{7}$$

After generating all iterations, the DE is then computes for each solution (chromosome) its fitness regard generating a high cluster quality in terms of inter-cluster dissimilarity and intra-cluster similarity using the above objective function Eq. 6.

2.5 Parameters Setup

This section discusses how to assign the DE's parameters run times which are: F, CR, and Population Size NP. The F parameter is a scale factor used to adjust the mutation process, and was set to 0.9 [7,10]. The CR is a user input parameter used to increase the diversity of vectors' parameters, and set to 0.5 [7,10]. The NP was set to 100 chromosomes per generation [7,10].

2.6 The Selected Similarity Measures: NGD

The similarity measure used is the Normalized Google Distance (NGD) [11], The NGD similarity is calculated between each two terms as in Equation 8:

$$sim_{NGD}(t_i, t_j) = exp(-NGD(t_i, t_j)) \tag{8}$$

Where

$$NGD(t_i, t_j) = \frac{\max\{\log(m_i), \log(m_j)\} - \log(m_{ij})}{\log(n) - \min\{\log(m_i), \log(m_j)\}} \tag{9}$$

m_i is the number of sentences including term t_i, m_{ij} refers to the number of sentences including terms t_i and t_j, and n is the total number of the document's sentences. Finally, to score the similarity between two sentences S_k and S_l is calculated using Eq. 10.

$$sim_{NGD}(S_k, S_l) = \frac{\sum_{t_i \in S_k} \sum_{t_j \in S_l} NGD(t_i, t_j)}{m_i m_j} \tag{10}$$

2.7 The Selected Features

We used the same features used by [13]: the Title feature, the Sentence Length feature, the Sentence Position feature, the Numerical Data feature and the Thematic Words feature. Table 1 shows how these features are calculated.

One of the differences of this work is using the feature-score concept to select the top and high relevance sentences to the document topic; unlike [1] which is used the sentence centroid to select the relevance sentences from each cluster to represent a summary sentences. In addition, scoring the sentence relevance in isolated way from other sentences makes the method is unable to capture the full advantage of the relationship between a sentence and other sentences in the document or a cluster [12].

Table 1. Selected features and their calculations [13]. Where S_i indexes the i^{th} sentence, and t is the total number of sentences on a given document.

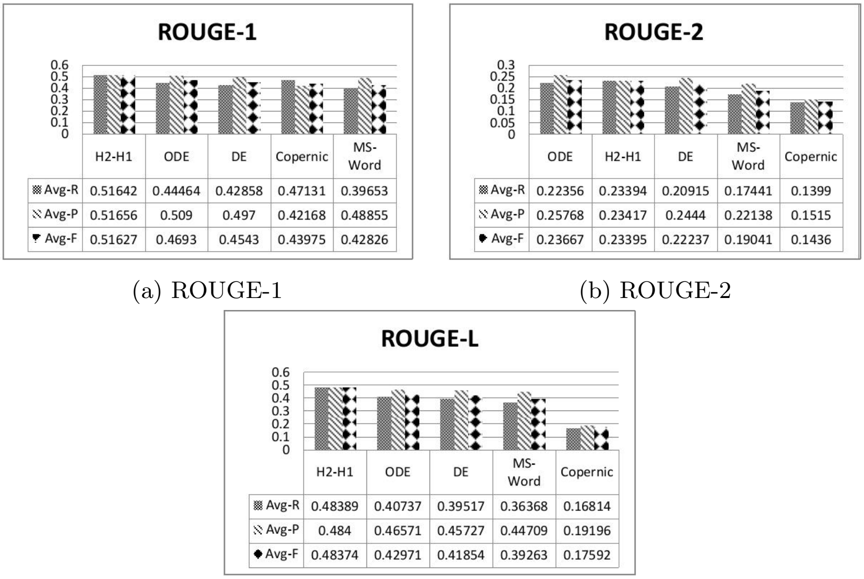
Feature	Calculation
Title	$\frac{\text{No. of}(S_i)\text{words matched title words}}{\text{No. of Title's words}}$
Sentence Length	$\frac{\text{No. of words in } S_i}{\text{No. of words in longest sentence}}$
Sentence Position	$\frac{(t-i)}{t}$
Numerical Data	$\frac{\text{No. of numerical data in } S_i}{\text{Sentence Length}}$
Thematic Words(TW)	$\frac{\text{No. of thematic words in } S_i}{\text{Max number of TW found in a sentence}}$

2.8 Dataset and Evaluation Measure

A set of 100 documents were collected from the standard summarization competition DUC 2002 [14]. The DUC submitted each news article document to two human experts to extract model summaries used for automatic system comparisons. In this experiment we assigned the first human summaries as reference summaries, while we compared the second human summaries against the first one (H2-H1) in order to evaluate the automatic systems performances with the human performance. The documents were preprocessed using the stop-words removal and words stemming. The ROUGE [8] is a standard automatic evaluation system in summarization. It includes a set of measures evaluate both single and multi-document summarization. ROUGE (1, 2, and L) are being used here as they are suitable measure when evaluate single document summarization [8]. The ROUGE toolkit produces recall, precision and F measures for each generated summary as well as a harmonic mean average for each one of these measures. It's worth mentioning that all evaluation scores extracted using ROUGE are statistically significant at the 95% confidence interval. For result analysis, we used the F-measure to evaluate the methods performance as it balances both recall and precision scores.

3 Experimental Result

Figure 2 shows and visualizes the extracted of sub-results rouge-1, 2, and L, respectively, using ROUGE summarization evaluation toolkit. The DE is the method proposed by [1] and ODE is our proposed Opposition DE method. Copernic and Ms-Word Summarizers are standard benchmark methods used widely in automatic text summarization comparison [15]. The results show that our proposed method outperformed all other methods at all evaluation measures. In addition, Figure 2.b shows that our proposed method was even outperformed



	H2-H1	ODE	DE	Copernic	MS-Word
Avg-R	0.51642	0.44464	0.42858	0.47131	0.39653
Avg-P	0.51656	0.509	0.497	0.42168	0.48855
Avg-F	0.51627	0.4693	0.4543	0.43975	0.42826

(a) ROUGE-1

	ODE	H2-H1	DE	MS-Word	Copernic
Avg-R	0.22356	0.23394	0.20915	0.17441	0.1399
Avg-P	0.25768	0.23417	0.2444	0.22138	0.1515
Avg-F	0.23667	0.23395	0.22237	0.19041	0.1436

(b) ROUGE-2

	H2-H1	ODE	DE	MS-Word	Copernic
Avg-R	0.48389	0.40737	0.39517	0.36368	0.16814
Avg-P	0.484	0.46571	0.45727	0.44709	0.19196
Avg-F	0.48374	0.42971	0.41854	0.39263	0.17592

(c) ROUGE-L

Fig. 2. Average Recall, Precision, and F-measure of all methods

the human performance at ROUGE-2. It's worth mentioning that all ROUGE evaluation scores extracted in this experiment are statistically significant using the confidence interval 95%.

4 Conclusion

In this paper, the machine learning OBL approach has been implemented to enhance the DE evolutionary search in terms of solution quality enhancement. The EAs save useful information during their runtime; hence the ML-OBL is used to extract this information and analysed it to study the search space behaviour. A number of related works were presented to enhance the DE evolutionary search, but all these works were limited only in learning DE in its real values mode only and tested on numerical functions test bed. In this work, we proposed to implement the OBL concept in order to generate more optimal solutions than the classical DE. Our presented work fitted the oppositional hypotheses as the much closest solutions to the complete ones as it was able to generate summaries more qualified than the classical DE. Our conclusion result showed that, enhancing the EA is a very important issue that should be considered as the performance of the algorithm is increased clearly than the random-based algorithm.

Acknowledgements. This work is supported by Ministry of Higher Education (MOHE) and Research Management Centre (RMC) at the Universiti

Teknologi Malaysia (UTM) under Research University Grant Category (VOT Q. J130000.7826.4F011). And deep thanks go to Mr. Ameer Tawfik Albahem for his valued time and hard Latex format assistance.

References

1. Alguliev, R.M., Aliguliyev, R.M.: Evolutionary Algorithm for Extractive Text Summarization. Intelligent Information Management 1(2), 128–138 (2009)
2. Jun, Z., Zhi-Hui, Z., Ying, L., Ni, C., Yue-Jiao, G., Jing-Hui, Z., Chung, H.S.H., Yun, L., Yu-Hui, S.: Evolutionary Computation Meets Machine Learning: A Survey. IEEE Computational Intelligence Magazine 6(4), 68–75 (2011)
3. Fattah, M.A., Ren, F.: GA, MR, FFNN, PNN and GMM based models for automatic text summarization. Computer Speech and Language 23(1), 126–144 (2009)
4. Binwahlan, M.S., Salim, N., Suanmali, L.: Fuzzy swarm diversity hybrid model for text summarization. Information Processing & Management 46(5), 571–588 (2010)
5. Alguliev, R.M., Aliguliyev, R.M., Isazade, N.R.: DESAMC+DocSum: Differential evolution with self-adaptive mutation and crossover parameters for multi-document summarization. Knowledge-Based Systems (2012)
6. Rahnamayan, S., Tizhoosh, H.R.: Differential Evolution Via Exploiting Opposite Populations. In: Tizhoosh, H.R., Ventresca, M. (eds.) Oppositional Concepts in Computational Intelligence. SCI, vol. 155, pp. 143–160. Springer, Heidelberg (2008)
7. Storn, R., Price, K.: Differential Evolution A Simple and Efficient Heuristic for Global Optimization over Continuous Spaces. J. of Global Optimization 11(4), 341–359 (1997)
8. Lin, C.-Y.: ROUGE: A Package for Automatic Evaluation of summaries. In: Proc. ACL Workshop on Text Summarization Branches Out, pp. 74–81. Association for Computational Linguistics (2004)
9. Ahandani, M.A., Alavi-Rad, H.: Opposition-based learning in the shuffled differential evolution algorithm. Soft Comput. 16(8), 1303–1337 (2012)
10. Vesterstrom, J., Thomsen, R.: A comparative study of differential evolution, particle swarm optimization, and evolutionary algorithms on numerical benchmark problems. In: Congress on Evolutionary Computation, CEC 2004 (2004)
11. Cilibrasi, R.L., Vitanyi, P.M.B.: The Google Similarity Distance. IEEE Transactions on Knowledge and Data Engineering 19(3), 370–383 (2007)
12. Shen, D., Sun, J.-T., Li, H., Yang, Q., Chen, Z.: Document summarization using conditional random fields. In: Proceedings of the 20th International Joint Conference on Artifical Intelligence, pp. 2862–2867. Morgan Kaufmann Publishers Inc., Hyderabad (2007)
13. Abuobieda, A., Salim, N., Eltayeb, R.A., bin Wahlan, M.S., Suanmali, L., Hamza, A.: Pseudo Genetic and Probablisitic-based Feature Selection Method for Extractive Single Document Summarization. Journal of Theoretical and Applied Information Technology (2011)
14. DUC, The Document Understanding Conference (DUC), http://duc.nist.gov
15. García-Hernández, R.A., Ledeneva, Y., Mendoza, G.M., Dominguez, Á.H., Chavez, J., Gelbukh, A., Fabela, J.L.T.: Comparing Commercial Tools and State-of-the-Art Methods for Generating Text Summaries. In: 2009 Eighth Mexican International Conference on Artificial Intelligence, pp. 92–96 (2009)

An Introduction to Ontology Based Structured Knowledge Base System: Knowledge Acquisition Module

Marek Krótkiewicz and Krystian Wojtkiewicz

Institute of Mathematics and Informatics,
University of Opole, Oleska 48, 45-048 Opole, Poland
{mkrotki,kwojt}@math.uni.opole.pl

Abstract. The following text presents the method of supplementing and verifying information stored in a framework system of the semantic knowledge base. The indicated method refers to the knowledge of ontological character, in other words to information about definitions of concepts and relationships among them. The aim of the method is the constant supplementing and verifying of the knowledge, and making more precise and detailed information about existing connections between concepts. The key aspect of the method is questions generating strictly dependent on the preconceived structure of stored knowledge.

Keywords: knowledge base, semantic knowledge, ontology, knowledge acquisition, knowledge verifying, knowledge processing, reliability of knowledge.

1 Introduction

The method described here was prepared for the needs of Semantic Knowledge Base (SKB) [1]. The main motivation of taking up this work was supplementing the engine of semantic knowledge base with a module of gaining the knowledge. The engine of the knowledge base, as a passive element, does not possess ability of acquisition, and it is not its task. The degree of complexity and specialization of knowledge structure in the system may prove simple methods of acquisition far from effective. In the case of relational or object-oriented data bases there is a possibility of analyzing structure of information and selecting appropriate methods of its acquisition in a way which remains in accordance with principles of integrity of these bases and semantics of information contained in them. In the case of the semantic knowledge base engine one cannot speak about established structure of information, since one of the types of stored information is the very structure. That is why it is dynamic and it is information itself, which highly complicates systematic, methodical approach to the issue of knowledge acquisition. Because of this, it has been assumed that the burden of supplementing knowledge and examining its integrity must be shifted to a specialized module. One of the system task is knowledge acquisition and the specialized module described here is to fulfill that task. It is its only task on which it focuses. For obvious reasons this task may be divided into many subtasks, such as e.g. verification of knowledge gained or examination of coherence, etc., which nevertheless is fully

A. Selamat et al. (Eds.): ACIIDS 2013, Part I, LNAI 7802, pp. 497–506, 2013.

contained within widely understood range of knowledge acquisition. It has also been assumed that SKB will treat a priority to the question of certainty of knowledge possessed by it, in relation to the amount of gathered information. An outcome of this approach is the complete resignation from exploring any uncertain, non-interactive knowledge sources e.g. Internet. The system focuses on interaction with the user-expert, who can answer questions, as well as, self-initiate introduction of information.

The aim of creation of such system was to elaborate software enabling the storage of knowledge. Neither scope of knowledge, which would be described, nor its field, nor the range of the use of the program was defined. Because of this, in its assumption, the system is a skeleton on the basis of which one can create specialized, dedicated AI solutions. One of the basic assumptions is that this system is constructed in module architecture, and each of the modules has a precisely defined role. For the purpose of this text we focused on two basic modules, i.e. the module of data storage (the engine of semantic knowledge base) and the module of active knowledge acquisition. To be able to present the method of the active knowledge acquisition one has to describe, at least to a certain extent, some of elements of the knowledge base engine. Its extensiveness, as well as, the degree of complexity make it impossible to present it fully. What is more it is the basis of a separate paper.

2 Basic Features of the Semantic Knowledge Base Engine

The engine of semantic knowledge base consists of data structures storing knowledge and mechanisms operating on it. Procedures connected with realization of low-level data operations and procedures of the higher level concerning knowledge processing will not be discussed here, since these are not included into the thematic main stream of this article. Data structure of the SKB has very strong influence on methods of knowledge acquisition, which will be shown while describing exemplary algorithms.

The structure has modular construction. The following modules are to be distinguished: Ontological Module (OM), Semantic Networks (SN)[7], Dimensions and Spaces Module, and a Linguistic Module. From active knowledge acquisition point of view the structure of the ontological module and linguistic module are crucial here. OM owes its name to its ability to store information that is recognized as ontological knowledge. Speaking precisely with the most important aspects of such knowledge type, i.e. concepts and relations among them. Part of the module focused on variety of relation classes, while the other stores information mainly about concepts, features and properties. The lexical module on the other hand plays the role of thesaurus that is used to provide names for concepts.

Concepts were divided into the following collections: CLASS, VALUE, FEATURE, ASSOCIATION etc., which is illustrated by figure 1.

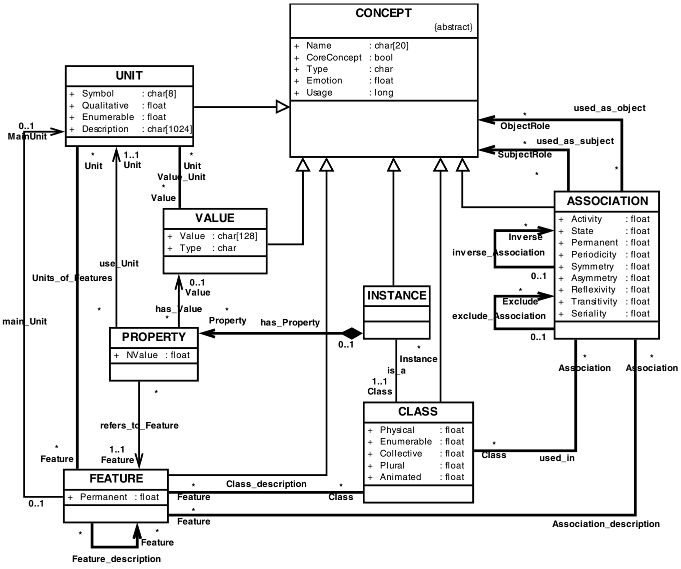


Fig. 1. Ontological Core class diagram

This division comes from semantic distinction of concepts classes used in SKB, while at the basis of this division lays object-oriented description of complexity. However, the object-oriented approach was not the only model used. The idea was also borrowed from a linguistic aspect of knowledge. Nevertheless, we gave up fully linguistic division into parts of speech and parts of sentence, since such division is a serious limitation of the ability to formulate knowledge. What is more, such division would have to be prepared for a defined natural language or linguistic group. Apart from the approaches mentioned here the module is strictly connected with the idea of frames, and, what flows from it, there is a possibility of easy information storage about structure: object, feature and value. This model has been slightly extended by adding possibility of defining the data types. Next important idea, which accompanied creation of the key module, was the idea of ontology. Ontology, as a specification of conceptualization [2][3], appears in the entire system. However, it was enriched by possibility of defining acceptable and unacceptable connections among concepts. This apparently small addition, in its intention, is the source of the power of the mechanism, which is essential for the verifying semantic correctness of knowledge provided. It comes from the fact that relations between concepts, as one of the ontological thoughts, can be verified in terms of sensibleness in a given space of concepts.

Another aspect in the description of knowledge structure in SKB is its multi-layer characteristics. On the lowest level of knowledge there are information about concepts and values of attributes describing them, leaving aside the layer of OODB, i.e. an engine securing storage of data on the lowest level. Next layer are binary connections between concepts, whose representatives can be such semantic relationships as taxonomy, part-whole, hyponymy, etc. Another layer consists of facts and rules, which, in an unconstrained way, describes complex information about existing connections among concepts. An important difference between this and the lower layer is complexity of connections described. The most universal way of storing knowledge are rules which, in their definition, have general character and allow the generating of facts and other rules.

3 The Levels of Minuteness of Information Detail and Possibility of Supplementing and Verifying Them

The most general and imprecise information is pure concept. Such concept possesses identifier which brings semantic information to a human subject. For example, if a concept will be named car, a man, possessing enormous ontological base in his mind, immediately positions this concept in a given space, associates it with other concept, automatically takes into account taxonomic connections, part-whole, synonyms, etc., and features. One identifier is enough to refer to voluminous knowledge. From the system point of view one has to take into account that, regardless of identifier applied, information about this concept is totally dependent on values of features and connections established earlier. In one word, if the new concept is introduced into the system, the system will gain knowledge in this way, only about the fact that "there is a term car." Such information is insufficient for concept introduction to SKB. However this is enough information for the lexical module. The system structure enforces separation of term(names) and concepts(meaning) what is shown on Figure 2[8].

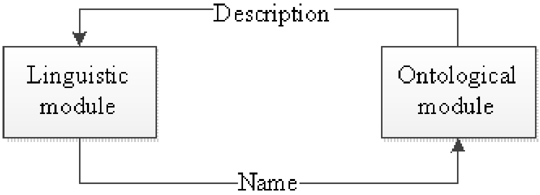


Fig. 2. Linguistic and ontological modules relation

Already on the level of an interface acquiring data (mainly the user interface) there exists visible separation of terms and concepts. It comes from the fact that the new concept, before it is defined, has to be ascribed to one of predefined collections according to the structure of the ontological core. In practice it means that after application of identifier (new concept), which was not known to system earlier, it will generate a question: "in which collection a new concept is to be classified?" In practice there is obviously a certain margin of choice, since collections are predefined. After receiving an answer the concept can be inscribed in the core. However, such information is very narrow, not to say almost of zero value, since apart from the existence of the concept, it says nothing about it. Because of this, there will take place automatic generation of another questions by the system, which will concern value of the attributes characteristic for a given collection. For the concept of car, classified in the collection CLASS, further detailed questions will ask for values of the following attributes: physical, enumerable, collective, plural and animated.

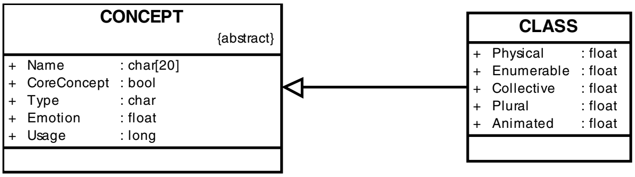


Fig. 3. CLASS collection, part of Ontological Core class diagram

Detailed description of attribute semantics will be the topic of wider monograph, which is still in the process of preparation.

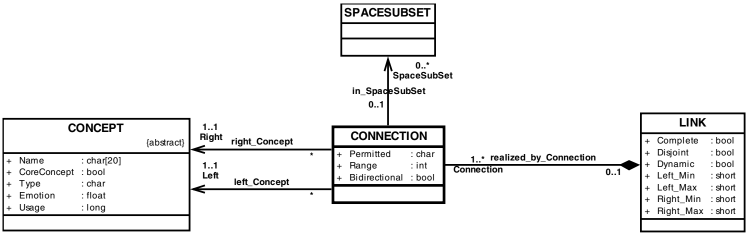


Fig. 4. CONNECTION collection, part of SSKBS class diagram

Side by side with value of features proper for a given collection, the system will supplement, partly in an automatic way, information about acceptable connections of a given concept with other concepts. At the very beginning system will try to use information from the lexical module to check possibility to ascribe connection to other concepts. Sometimes the name "label" may be changed or rather new names may be connected to specific concept. It will perform the task on the basis of analysis of the sentence introduced, obviously after its earlier grammatical analysis. The question of communication with SKB, thus of grammar and semantics of such language, is also broad issue and goes beyond the thematic scope of this article.

Moreover it is not a key issue for the aspect of automatic knowledge acquisition. In this moment it is possible for the user to supplement information about acceptable connections between the new concept and concepts already defined in the system. However, from a practical point of view supplementing such information would be extremely toilsome for a man, since the number of acceptable connections among concepts is enormous. Therefore the burden of automatic supplementing of such information falls on the system. As we have already said, such information is essential from the verification of sentences semantic consistence point of view. A separate issue arises while defining relation with other concepts in spite of e.g. taxonomy, connection part-whole, etc., and other non-predefined, free connections. As contrasted with binary connections which were described as acceptable connections of two concepts, this module describes acceptability of ternary connections. Semantics of this information points to the fact that, if there is trinity of concepts A-B-C, it means that concepts A and C are and can be related to B. At the same time it is accepted, for the sake of optimization, that A and C can be not only concepts but also sets of concepts connected by a concrete relation. One has to pay attention to the fact that similarly to the case of binary connections, there is possibility of recognition if such connection exists or is only possible. It opens wide possibilities for defining new dependencies. For example it widens the concept of the object model by including taxonomic connections which express not only the fact of inheritance but also the fact of acceptability of inheritance as a specific option. Semantic of attributes of such connections is complex from the mathematic point of view, as well as algebra of connections in the object model. It will not be described here. There are many other properties that can be generated on the basis of predefined types of ternary connections linked with attributes of these connections, and, what is more, the infinite set of undefined relations give possibility of description of complexity of the world. What deserves attention is the fact that so far we have left aside, practically completely, description of the essential element of SKB thanks to which it is possible to define concepts, their binary and ternary connections for a given space. It enables separation of elements mentioned above, and in this way creation of many worlds (spaces), in which concepts may exists parallel to each other. Thanks to it, spaces can penetrate each other and create common sub-spaces. In connection with the mechanisms presented above, the idea of objectivity widened by additional functionalities, frames, structure of relations, effect of synergy, open the wide range of possibilities of describing complexity of the surrounding world [5]. That is why the aspect of the automatic supplementing and verifying of knowledge becomes extremely important, since control of such complexity exceeds capabilities not only of one man but also of a group of experts in a given field[4]. One has to add that the described ternary connections in fact are multiple thanks to the mechanism of defining arguments of relationships in the form of sets, which after simplification of description was not exposed above.

The areas of activity concerning automatic acquisition of knowledge, described above, were connected mainly with the ontological module. As we have already mentioned, this level of knowledge is essential for verification of correctness of knowledge stored in rules and facts [6]. Automatic acquisition of knowledge concerns also facts and rules, that are stored in the semantic network module. It is a more complex issue and it is impossible to describe the way of its realization without presentation of language in which rules and facts are stored. Grammar and semantics of such language is the basis on which algorithms, employed for the generating of supplementing and verifying questions, are built. It is obvious, therefore, that in the first place one has to define precisely the language of the system which in fact goes beyond the limits of this paper. In conclusion one has to say that the mechanisms of this module will not be presented here, since it would force us to refer to details of structure, algorithms and methods, which will not be contained in this article. Therefore in the further part we focus on description of characteristic algorithms based on the ontological core and serving for the generating of key questions and its use of knowledge for verification of semantic inner coherence, which is already present in the system or will be introduced in future.

4 General Form of Active Knowledge Acquisition Algorithm

The basis of knowledge acquisition process is interactive communication with an expert. Introducing an utterance to the SKB it is necessary to assign terms to appropriate collections. After that these collections are searched for concepts that could be used by the term. If a concept is not found, it means that it is a new concept and the procedure of its introduction to the ontological core should be started. This procedure consists on recognition of values of features characteristic for a given collection. These features will allow processing interiorly utterances, but mainly they are to differentiate concepts. One has to add that attributes of concepts should be filled in such a way as to ascribe each a certain value. Lack of answer about any attribute in majority of cases will be significant weakness and, what is connected with it, it will make knowledge processing more difficult. After introduction of a concept, together with values of attributes describing it, into the base, it should be ensured that this concept will be connected with other concepts. Connections may be, as we have said, both possible and forbidden. Depending on a concept we are dealing with, and more precisely on collection, one has to supplement other sets of connections. For example, a particular case is the concept of INSTANCE collection. When it appears, one has to necessarily recognize a class to which it belongs, i.e. connection INSTANCE-CLASS. Assigning the class, we get automatically the list FEATURE objects, and what comes from it, we are able to ascribe values, i.e. establish the connection PROPERTY-INSTANCE. Thanks to this mechanism the expert receives a number of questions about the PROPERTY list, which has to be provided for a given INSTANCE. Connection INSTANCE-CLASS, and further CLASS-FEATURE makes the mechanism, since FEATURE for CLASS is the same as PROPERTY for INSTANCE. In the other cases e.g. objects of collection CLASS, FEATURE, ASSOCIATION, in accordance with a given scheme, one has to supplement connections with remaining exemplars of the given collection. Thus for example, appearance of a new concept from the collection ASSOCIATION will generate a question about connection with the collection CLASS and FEATURE. Going further the concept from the collection CLASS will generate a question about connection with the collection ASSOCIATION and FEATURE, in the case of FEATURE with ASSOCIATION and CLASS. As we see, there occurs recursion of data, and in fact, also recursion of algorithm of generating questions. This process in a general case is obviously infinite, which may be expressed by one sentence: the more we know, the more questions there appear. This illustrates perfectly provided algorithm. Theoretically, if one was to introduce into the system only one concept, chosen at random, regardless of its minuteness or generality, the above method allows system generating of all the concepts in a given space familiar to the expert.

The process described here, in fact, any set of its motifs, can be, and more precisely, is interrupted, which is a result of decision of the expert or of desistance of further scrutiny of the connection graph. One has to emphasize that the above process concerned only binary connections. In the case of ternary and multiple connections the same method, in terms of idea, is applied. However, algorithm in this moment is slightly more complicated for obvious reasons. The effect is, however, similar, namely there occurs avalanche generation of questions about connections, which results in extremely dynamic growth of

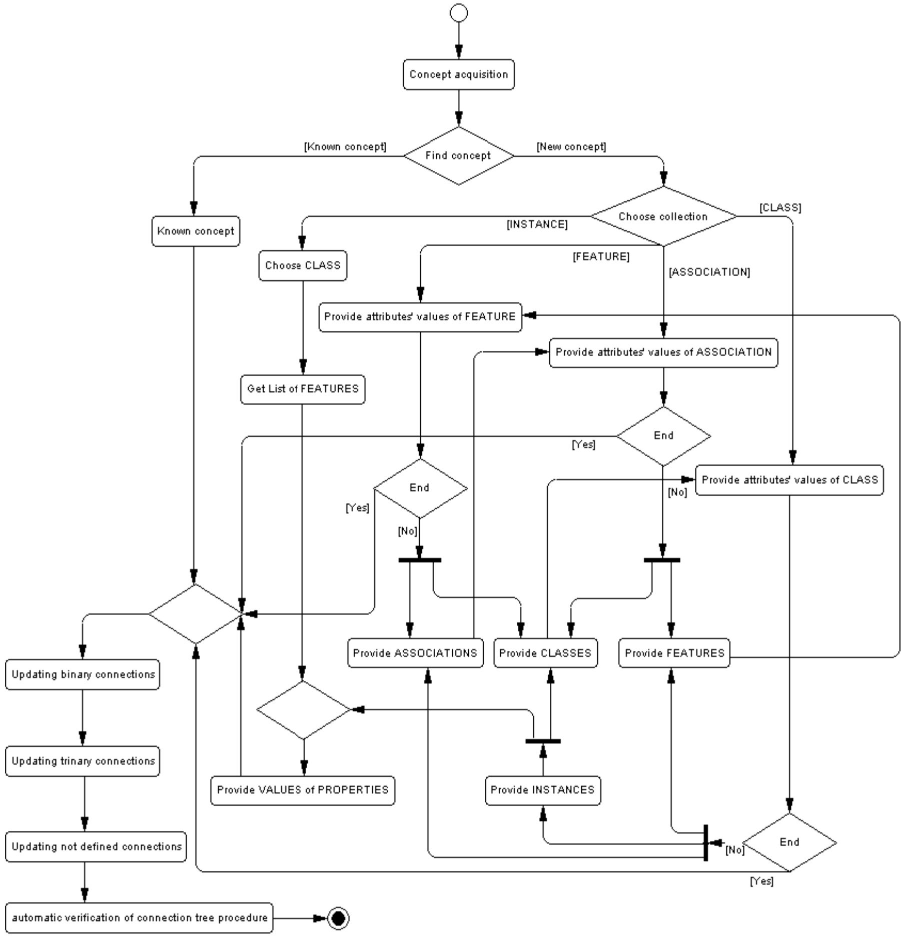

Concept acquisition
[Known concept]
Find concept
[New concept]
Known concept
Choose collection
[INSTANCE]
[CLASS]
Choose CLASS
[FEATURE]
[ASSOCIATION]
Provide attributes' values of FEATURE
Provide attributes' values of ASSOCIATION
Get List of FEATURES
End
[Yes]
Provide attributes' values of CLASS
[No]
End
[Yes]
[No]
Updating binary connections
Provide ASSOCIATIONS
Provide CLASSES
Provide FEATURES
Updating trinary connections
Provide VALUES of PROPERTIES
Provide INSTANCES
Updating not defined connections
[No]
End
[Yes]
automatic verification of connection tree procedure

Fig. 5. Knowledge acquisition algorithm

knowledge in a given field. The process of generating multiple connections is semantic and logical consequence of binary connections. There exist concrete rules of semantic coherence, quite obvious, if we take into account that each multiple connection can be recorded as a set of binary connections. Because of this, there exists also dependency in terms of generating such connections. From the generating of questions point of view there is a certain kind of positive informative feedback between generator of n-nary and binary connections. It means that not only binary connections are necessary for n-nary ones, but also potential new n-nary connections will generate binary ones. Recapitulating, one had to emphasize the fact that the method of generating questions, presented here, that has to provide system with systematic supplementing of knowledge, operates on the basis of the system of quite complex data structure. This structure is multi-leveled, hybrid in terms of knowledge storing models, from the definitely object approach, through frames, linguistic approach and a number of additional mechanisms widening these models. It results in necessity of exceptionally concise description of the idea of generating questions, since to understand fully the mechanisms described one would have to refer to an extremely rich description of structures and algorithms of the very system.

5 Conclusion

Acquisition of knowledge is one of the most important aspects of its existence. The system, in the case described above, exists and undertakes actions only for one purpose. The sense of its existence starts and ends with development of the knowledge base. Other actions of AI are only derivatives or flow from it directly or indirectly. The very issue is, from the point of AI experts view, one of the most significant, was discussed in numerous sources. Many of presently very popular approaches focuses on a large-scale knowledge acquisition. In this task Internet become very attractive. They aim at searching Internet and treating it as a knowledge base. According to the authors of this article, Internet is not a knowledge base, on which autonomous systems can rely. It can be a certain set of information, popularly and unjustly named a knowledge base, however, it is only for such advanced systems as man himself presents. The authors accepted different assumptions and decided to focus not on amount of data, and in fact knowledge, but on its coherence, verifiability and certainty. The reason underlying such approach is another assumption, that more damages and problems appears not because of the lack knowledge, but because of false and in the best case uncertain one.

Further research over the system have many directions. One of them is user's interface, which is strictly connected with the aspect of acquisition of knowledge. In the above article the description of semantic network module, in which there concentrates the main task of storing complex information, was completely ignored. Complexity of these information, and more precisely of their structure connected with complexity of communication language, necessitates exceptionally precise and strictly well-fitting user's interface enabling as proficient as possible exchange of statements. This task is complex and because of the volume of the project it creates numerous problems, since its various modules are found on various stages of development, which will be the topic of future publications.

References

1. Krótkiewicz, M., Wojtkiewicz, K.: Conceptual ontological object knowledge base and language. In: Kurzyński, M., Puchała, E., Woźniak, M., Żołnierek, A. (eds.) Computer Recognition Systems Proceedings of the 4th International Conference on Computer Recognition, pp. 227–234. Springer, Berlin (2005)
2. Gruber, T.R.: A Translation Approach to Portable Ontology Specifications. Knowledge Acquisition 5(2), 199–220 (1993)
3. Gruber, T.R.: Toward principles for the design of ontologies used for knowledge sharing. Originally. In: Guarino, N., Poli, R. (eds.) International Workshop on Formal Ontology, Padova, Italy (revised August 1993); Published in International Journal of Human-Computer Studies 43(5-6), 907–928 (1993, November/December 1995)
4. Kulikowski, J.L.: Data Quality Assessment. In: Encyclopedia of Database Technologies and Applications, pp. 116–120 (2005)

5. Duong, T.H., Nguyen, N.T., Jo, G.: A Hybrid Method for Integrating Multiple Ontologies. Cybernetics and Systems 40(2), 123–145 (2009)
6. Krótkiewicz, M., Wojtkiewicz, K.: Knowledge Acquisition in Conceptual Ontological Artificial Intelligence System. In: Hippe, Z.S., Kulikowski, J.L. (eds.) Human-Computer Systems Interaction. AISC, vol. 60, pp. 29–37. Springer, Heidelberg (2009)
7. Hendrix, G.G.: Expanding the utility of semantic networks through partitioning. In: Proceedings of the 4th International Joint Conference on Artificial Intelligence, Tbilisi, Georgia, USSR, pp. 115–121 (1975)
8. Nirenburg, S., Raskin, V.: Ontological Semantics. The MIT Press, Cambridge (2004)

Comparison of Gene Co-expression Networks and Bayesian Networks

Saurabh Nagrecha[1], Pawan J. Lingras[2], and Nitesh V. Chawla[1]

[1] Department of Computer Science and Engineering,
University of Notre Dame, Indiana 46556, USA
[2] Department of Mathematics and Computing Science, Saint Mary's University,
Halifax, Nova Scotia, Canada B3H 3C3

Abstract. Inferring genetic networks is of great importance in unlocking gene behaviour, which in turn provides solutions for drug testing, disease resistance, and many other applications. Dynamic network models provide room for handling noisy or missing prelearned data. This paper discusses how Dynamic Bayesian Networks compare against coexpression networks as discussed by Zhang and Horvath [1]. These shall be tested out on the genes of yeast *Saccharomyces cerevisiae.* A method is then proposed to get the best out of the strengths of both models, namely, the causality inference from Bayesian networks and the scoring method from a modified version of Zhang and Horvath's method.

1 Introduction

Biological processes, and by extension life, emerge from processes at the most basic level of the cellular structure- genes and proteins. A highly structured system of networks is responsible for information flow through the cell.

The central dogma of biology suggests mechanisms of information transfer in biological networks. This requires for us to consider genes, proteins, and their mutual interactions. DNA replication, transcription and translation are a few of these processes via which information is transferred. Gene coexpression analysis aims to provide increasingly reliable interaction models of biological systems. We restrict our model to that of a system of genes interacting with each other via expression. The nodes represent the individual genes, edges represent interactions within the system. These networks may be directed or undirected, cyclic or acyclic.

Gene expression studies usually start with microarray experiments where the expression levels of thousands of genes can simultaneously be measured. Microarray gene expression experiments are done with specimens of known heritage. These are exposed to a controlled environment with variables like nutrition, illumination, presence of various concentration of drugs. These experiments typically generate large matrices of gene expression levels. This data is usually quite noisy and may have missing values.

This data is then used to answer questions about regulatory mechanisms of gene expression. The authors demonstrate the performance of Bayesian Networks

A. Selamat et al. (Eds.): ACIIDS 2013, Part I, LNAI 7802, pp. 507–516, 2013.

as compared to coexpression networks, validated against curated gene interaction data.

2 Literature Review

A number of sophisticated methods which answer specific questions have been developed and proposed through the past two decades. Groundbreaking work by Spellman *et al.* in 1998 [2] on yeast genes using microarray hybridization techniques opened the field of systems biology and made it possible to perform scalable operations on genetic datasets. Applications ranging from the humble yeast to the Human Genome Project ultimately aim to create a "Rosetta Stone" to decipher the mystery that biological systems pose [3].

Approaches using Boolean Networks [4][5], and the next logical step Artificial Neural Networks have been proposed [6]. Methods using independent component analysis, and then self organized maps were used by Dragomir [7] were employed to solve the problem of class discovery.

The model used here builds on the approach by Murphy and Mian in [8]. Their method deals with Bayesian (belief) Networks as discussed in [9]. It unifies and generalizes models of boolean networks, Hidden Markov Models, and other widely accepted models. Boolean networks and Hidden Markov Models can be shown to be interconvertible with suitable assumptions of an intermediate state vector. Markov chains come associated with an inherent transition matrix (T) and if $T(i, j) = 0$, then this means that the system cannot make the transition from state i to state j. This kind of representation is unsuitable for sparse, discrete models- the kind we're considering here. So, we do not consider Boolean Networks or HMMs.

The use some or all of the aforementioned methods in yeast genes (those of the *Saccharomyces cerevisiae*) is of specific interest because it is fully sequenced, and widely researched. Bilu and Linial's [10] work proposes a hierarchical clustering through the metric "BLAST" which is a measure of similarity in genes. A functional prediction is then performed so as to validate the clustered genes.

Yeast genes are studied using Bayesian Networks by Friedman, *et al* in [11]. This Bayesian Network is put through a validation of known experimental results. The procedure is suited to cell cycle expressions and is thus of direct importance to our proposed method.

The system of coexpression networks inferred via a modification of the methods by Zhang and Horvath [1] for each timeframe is considered as an instance in a Markov Chain. This is then collapsed into a Bayesian Network (as justified above) using the networks discussed by Friedman *et al* in [11].

3 Study Data and Experimental Design

3.1 Study Data

As a consequence of the extensive nature of DNA microarray experiments, a "genomic" viewpoint on gene expression is provided. Data from microarray experiments on *Saccharomyces cerevisiae* by Spellman *et al.* is used here to demonstrate the methods proposed. This dataset contains 76 gene expression measurements of mRNA levels of 6177 *S. cerevisiae* ORFs. This data represents six time series under different cell cycle synchronization methods. According to Spellman *et al.* about 800 genes exist whose expression varies over different stages of the cell cycle. This data contains about 6% missing values which shall be dealt with slightly differently in the methods discussed below.

This data contains real values from the experiments. Usually, this is discretized for most purposes into 3 categories: *underexpressed* (-1) *baseline/normal* (0) and, *overexpressed* (1), depending on whether the gene is expressed lower than, similar to, or greater than the control, respectively. The thresholds for such discretization may be arrived at by setting the average from across the experimental data or from other independent experiments.

3.2 Coexpression Networks

Coexpression networks treat each gene as an individual node and connections between two such nodes depict the nature of interaction between the two genes. These interactions depend on the complexity of the model chosen. For instance, one could choose binary edges to denote presence (edge weight=1) or absence (edge weight=0) of interaction. Softer thresholding methods enable us to define weighted edges in the coexpression network. Adjacency functions which return such weights need to be defined accordingly. The parameters for these are sought using biologically motivated criteria, viz. the scale-free topology criterion[12,13].

Measures of Gene Similarity. Data is often taken in the form of raw expression levels where missing data usually results in loss of valuable information. A modified version of the data is used in this case instead. Exploiting the temporal nature of the time-series data, a noise eliminating curve-fit is implemented to take care of missing values and to smoothen out noisy kinks. This results ina relatively noiseless and more reliable correlation score. The similarity between each pair of genes is denoted by the measure s_{ij}. The absolute value of the Pearson correlation coefficient $s_{ij} = |cor(i,j)|$, or a shifted-and-scaled version $s_{ij} = \frac{1+cor(i,j)}{2}$ of it are often used here. The aim is to arrive at a similarity measure lying between 0 and 1. The similarity matrix thus arrived at, is denoted by $S = [s_{ij}]$

The Adjacency Function. To transform the similarity matrix into an adjacency matrix, the adjacency function is applied. The choice of the adjacency function decides whether we have soft (resulting in a weighted network) or hard thresholding (resulting in an unweighted network). The adjacency function is required to be a monotonically increasing function which maps the interval [0,1] into [0,1]. Hard thresholding for example works as below:

$$a_{ij} = signum(s_{ij}, \tau) \equiv \begin{cases} 1 \text{ if } s_{ij} \geq \tau \\ 0 \text{ if } s_{ij} < \tau \end{cases}$$

Soft thresholding is implemented so as to mitigate the loss of information incurred by hard thresholding. Two types of soft thresholding methods are often used: The sigmoid function

$$a_{ij} = sigmoid(s_{ij}, \alpha, \tau 0) \equiv \frac{1}{1+e^{-\alpha(s_{ij}-\tau 0)}}$$

and the power adjacency function

$$a_{ij} = power(s_{ij}, \beta) \equiv |s_{ij}|^{\beta}$$

Opinion on methods of estimating parameters for these functions varies widely. Methods suggesting the usage of p-value instead of the correlation coefficient in order to impose a hard threshold are commonly used. For soft thresholding methods, a detailed treatment using scale-free topology criteria is shown in [1]

Node Similarity/Dissimilarity. The coexpression analysis aims to identify tightly connected subsets of nodes. Out of many dissimilarity measures defined by authors, the toplogical overlap of two nodes [14] was shown to be useful in biological networks. For unweighted networks, the measure can be shown as below:

$$\omega_{ij} = \frac{l_{ij}+a_{ij}}{min\{k_i,k_j\}+1-a_{ij}}$$

where $l_{ij} = \sum a_{iu}a_{uj}$ and $k_i = \sum a_{iu}$. This may as well be extended to weighted networks. Here, in the case of $\omega_{ij} = 1$, the node with the lesser degree satisfies two conditions: 1) all of its neighbors are also neighbors of the other node and 2) it is connected to the other node. On the contrary, $\omega_{ij} = 0$, if i and j are unconnected and the two nodes do not share any neighbors. The topological overlap matrix is thus $\Omega = |\omega_{ij}|$ and is non-negative and symmetric. The dissimilarity measure is simply $d^{\omega}_{ij} = 1 - \omega_{ij}$. This matrix is the one which leads to distinctly clustered gene modules.

3.3 Bayesian Networks

Friedman *et al.*'s method treats the data with no prior assumptions of biological knowledge. It initially treats the measurements as independent samples from a distribution, ignoring the temporal aspect of the measurement. This is compensated by introducing an additional variable to denote the cell cycle phase. This variable is of key significance in all the networks learned and is forced to be a root in all the networks learned. This translates to the expression levels of the genes being dependent on the cell cycle phase.

Mathematical Formalism. A Bayesian Network is a representation of a joint probability distribution, comprising two components: the topological component G is a directed acyclic graph (DAG) whose vertices correspond to the random variables $X_1, ..., X_n$ and the second being Θ , the conditional distribution for each variable, given its parents in G. These components combined form a unique distribution in the space of the random variables $X_1, ..., X_n$. In association with the *$Markov\ Assumption$*, which states that each variable X_i is independent of its nondescendants, given its parents in G, the graph is a compact representation of the joint probability distribution, thus economizing on the number of parameters. The chain rule of probabilities and properties of conditional independencies help us decompose this into the $productform$ as below:

$$P(X_1, ..., X_n) = \prod P(X_i|\mathbf{Pa}^G(X_i))$$

where $\mathbf{Pa}^G(X_i)$ is the set of parents of X_i in G. The conditional distributions $P(X_i|\mathbf{Pa}^G(X_i))$ for each variable X_i are denoted by parameters specified in θ.

In specifying the conditional probability distributions, it is usual for one to represent the input random variables as continuous, discrete or mixed (in keeping with our initial mention of how the expression data is represented and subsequently interpreted). Continuous variables are usually represented using multivariate linear Gaussian distributions as $P(X|u_1, ..., u_k \sim N(a_0 + \sum a_i \cdot u_i, \sigma^2))$. Here the normally distributed random variable X's mean linearly depends on the values of its parents. If all the variables have similar Gaussian conditional distributions, then the joint distribution is a multivariate Gaussian. Discrete variables can be represented by multinomial distributions. This makes the free parameters exponential in the number of parents. Hybrid networks contain a mixture of continuous and discrete variables. These are of little relevance here and discussed in greater detail in Friedman *et al.*'s paper.

Learning Bayesian Networks. Learning a Bayesian Network can be stated as a problem as follows. Given a training set D=$\{X_1, ..., X_n\}$ of independent instances of $\mathcal{X}$, find a network $B = < G, \Theta >$ that best matches D. The problem is that of an optimization in the space of directed acyclic graphs in. The number of such graphs is superexponential in the number of variables involved.

An algorithm by Friedman, Nachman and Pe'er, called the *Sparse Candidate* algorithm is an efficient search procedure which focuses on certain relevant regions of the search space. We can identify a few key candidate parent node genes based on local statistics (like correlation) and then restrict the search to only those networks where these candidates are the parent nodes, thus slicing down the search space considerably and quickly. Possible over-restriction of the search space can be dodged by implementing an iterative search for the optimum set of initial candidates. The descendents are then found after ascertaining these candidates [8].

3.4 Experimental Design

The models arrived at using the above discussed methods are compared based on their performance with respect to biological data that has been experimentally validated. This provides a comparative study of how robust each method is, both individually, as well as when compared to each other. The performance may be based on the models' ability to identify important topological structures and causal patterns. Some of which are described below:

Dominant Genes. Genes directly involved in initiation and control of cellular cycles can be perceived as nodes of prime importance. Accuracy in predicting properties of such nodes is a cursory measure of robustness of the model.

Prominent Motifs. Extending the previous argument further topologically, it could be said that not just key nodes, but some motifs play an important role in cellular processes. The degree of isomorphism of the cell cycles (which would appear as subgraphs) identified by the methods with respect to the original motifs in the pre-validated data could be considered a more sophisticated extension of the above metric.

Markov Relations. Functional relations which make biological sense can be inferred using Markov chains as the criteria of identification. Functionally related genes which can be represented as Markov chains are an important feature. High confidence Markov relations have been known to concur with experimental validation. More interestingly, among high confidence Markov chains, one can often find conditional independence i.e. a group of highly correlated genes. [11]

Due to the extensive research performed on *Saccharomyces cerevisiae*, such data is readily available in the works of Spellman, etc. Thus the two methods are pitted head to head against each other. The validation is carried out for the most prominent genes in the organism and subsequent inferences are made.

4 Implementation

A sample data-set consisting of 12 key genes 77 states of observation (with no missing values) was used to test out most of the methods. This dataset required no cleanup due to the choice of the genes. The main network inference was carried out with the data of 6178 genes with 77 different observation states. This data contains numerous missing values.

4.1 Data Cleaning

In inferring a Bayesian Network out of such an incomplete database as the *S.cerevisiae* expression data, one is presented with a choice between ignoring the missing values and adapting dynamic models. A compromise between these two methods is chosen by using a modified dataset which now comprises of 58 observation instances instead of the original 77, the upper quartile of the missing values being omitted. This subset is further trimmed to provide a complete dataset without missing data for any of the constituent genes. This elminates the requirement to assume the presence of any further fictitious nodes and simplifies the complexity of the problem significantly.

Prevalidated interaction data is obtained from the "Saccharomyces Genome Database". The data of relevance here is in the form of arcs constituting respective gene names. This contains a few repeated interactions, which are eliminated.

4.2 Bayesian Networks

The Bayesian networks can be formed using two approaches- score based structure learning algorithms like hill-climbing and Tabu search, and using constraint based structure learning algorithms. Constraint-based algorithms offer flexibility in setting thresholds for false positive, Type I statistical errors; but at the same time their execution time is greater than that of score based structure learning algorithms. Under similar values of α, the various constraint based algorithms seem to perform equally in terms of predicting arcs. A precision-recall analysis is done by observing the effect of varying α.

4.3 Coexpression Networks

The coexpression networks are formed using Pearson Correlation Coefficient or Mutual Information based scores using a dynamically derived β (not to be confused with the Type II statistical error). This method, unlike Bayesian Network inference, is robust to missing data (as long as the data stays within statistically significant margins). Here we adhere to the Pearson Correlation Coefficient based methods and implement hard thresholding using random resampling methods by bootstrapping the data. In doing so, we assume a Gaussian-like distribution of the expression values (which it does resemble closely). The final clusters themselves are of little significance for this particular analysis and we focus merely on the edges obtained.

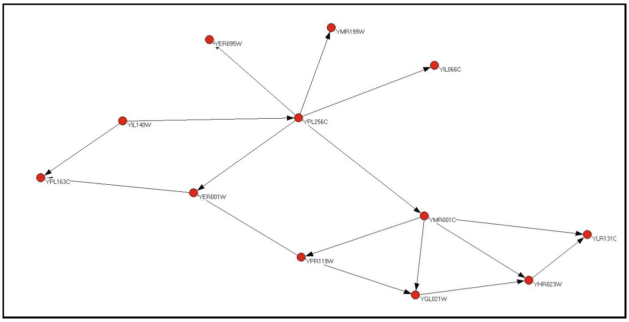


Fig. 1. The network formed out of the 12 genes considered. This Bayesian Network reflects the causal relationships exhibited in the gene cycles.

5 Inferences and Conclusions

5.1 Preliminary Inferences

Networks made using the toy dataset using coexpression analysis and Bayesian Network models concur in their predictions and with the validated dataset. This may be attributed to the small sample size of the subset chosen. The network exhibited is as per Fig. 1.

The final network consisting of 2361 genes and 7182 edges are displayed in the Pajek visualization as shown in Fig.2.

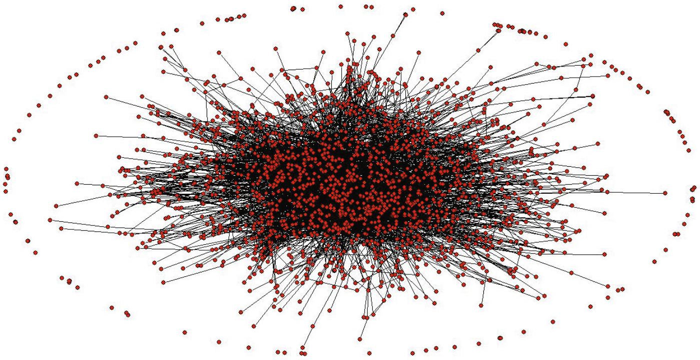


Fig. 2. Network comprising 2361 genes of *S.cerevisiae*. A Pajek visualization of the coexpression network.

Results of the Bayesian networks reveal the following key Markov relations in particular as per Table 2.

Table 1. Performance statistics for the different models. Note that the terms "precision" and "recall" are used in the context of the Bayesian Network being the relevant data and the coexpression network is the retrieved data.

Set A	*Set B*	*Precision*	*Recall*
Validated Data	Bayesian Network	0.95	0.67
Validated Data	Coexpression Network	0.62	0.53
Bayesian Network	Coexpression Network	0.83	0.77

5.2 Validation and Inferences

We perform a precision-recall analysis between the 3 set of arcs: the prevalidated data, the Bayesian Network, the coexpression network. (as shown in Table 1) The Bayesian Network is seen to be a closer estimater of gene interactions than the coexpression network due to superior precision and relatively higher recall. Relations that are prominently expressed in the data appear in all 3 models. The biological interpretations of the interactions are either spatial or functional in nature. Despite this, few of the high confidence functional interactions predicted may be considered false positives if arrived at using a Gaussian model, as it uses correlation values. This problem does not arise in the multinomial model, whose salient results are as per Table 2.

Table 2. List of Top Markov Relations, Multinomial Experiment

Confidence	*Gene 1*	*Gene 2*	*Notes*
1.0	YKL163W-PIR3	YKL164C-PIR1	Close locality on chromosome
0.985	PRY2	YKR012C	Close locality on chromosome
0.985	MCD1	MSH6	Both bind to DNA during mitosis
0.98	PHO11	PHO12	Both nearly identical acid phosphatases
0.975	HHT1	HTB1	Both are histones
0.97	HTB2	HTA1	Both are histones
0.94	YNL057W	YNL058C	Close locality on chromosome
0.94	YHR143W	CTS1	Homolog to EGT2 cell wall control, both involved in cytokinesis
0.92	YOR263C	YOR264W	Close locality on chromosome
0.91	YGR086	SIC1	Homolog to mammalian nuclear ran protein, both involved in nuclear function
0.9	FAR1	ASH1	Both part of a mating type switch, expression uncorrelated
0.89	CLN2	SVS1	Function of SVS1 unknown
0.88	YDR033W	NCE2	Homolog to transmembrame proteins suggest both involved in protein secretion
0.86	STE2	MFA2	A mating factor and receptor
0.85	HHF1	HHF2	Both are histones
0.85	MET10	ECM17	Both are sulte reductases
0.85	CDC9	RAD27	Both participate in Okazaki fragment processing

5.3 Conclusions

In this paper, a comparison has been made between Bayesian Network and coexpression networks on the basis of performance in predicting the structure of the expression network of the genome for baker's yeast. This is done without any prior knowledge of biology involved; in fact, biologically viable and plausible interactions stem out of the predicted models. A more throughly biologically supervised global topological treatment has been discarded in favor of learning the finer interaction structure. Evidently, Bayesian networks emerge as a more informative tool to determine the causal structure of such interactions.

References

1. Zhang, B., Horvath, S., et al.: A general framework for weighted gene co-expression network analysis. Statistical Applications in Genetics and Molecular Biology 4(1), 1128 (2005)
2. Spellman, P., Sherlock, G., Zhang, M., Iyer, V., Anders, K., Eisen, M., Brown, P., Botstein, D., Futcher, B.: Comprehensive identification of cell cycle–regulated genes of the yeast saccharomyces cerevisiae by microarray hybridization. Molecular Biology of the Cell 9(12), 3273–3297 (1998)
3. Ideker, T., Galitski, T., Hood, L.: A new approach to decoding life: systems biology. Annual Review of Genomics and Human Genetics 2(1), 343–372 (2001)

4. Akutsu, T., Miyano, S., Kuhara, S., et al.: Identification of genetic networks from a small number of gene expression patterns under the boolean network model. In: Pacific Symposium on Biocomputing, vol. 4, pp. 17–28. World Scientific, Maui (1999)
5. Hakamada, K., Hanai, T., Honda, H., Kobayashi, T.: A preprocessing method for inferring genetic interaction from gene expression data using boolean algorithm. Journal of Bioscience and Bioengineering 98(6), 457–463 (2004)
6. Weaver, D., Workman, C., Stormo, G., et al.: Modeling regulatory networks with weight matrices. In: Pacific Symposium on Biocomputing, vol. 4, pp. 112–123. World Scientific, Maui (1999)
7. Dragomir, A., Mavroudi, S., Bezerianos, A.: Som-based class discovery exploring the ica-reduced features of microarray expression profiles. Comparative and Functional Genomics 5(8), 596–616 (2005)
8. Murphy, K., Mian, S., et al.: Modelling gene expression data using dynamic bayesian networks. Technical report, Computer Science Division, University of California, Berkeley, CA (1999)
9. Pearl, J.: Bayesian networks (2011)
10. Bilu, Y., Linial, M.: The advantage of functional prediction based on clustering of yeast genes and its correlation with non-sequence based classifications. Journal of Computational Biology 9(2), 193–210 (2002)
11. Friedman, N., Linial, M., Nachman, I., Pe'er, D.: Using bayesian networks to analyze expression data. Journal of Computational Biology 7(3-4), 601–620 (2000)
12. Barabási, A., Oltvai, Z.: Network biology: understanding the cell's functional organization. Nature Reviews Genetics 5(2), 101–113 (2004)
13. Rzhetsky, A., Gomez, S.: Birth of scale-free molecular networks and the number of distinct dna and protein domains per genome. Bioinformatics 17(10), 988–996 (2001)
14. Ravasz, E., Somera, A., Mongru, D., Oltvai, Z., Barabási, A.: Hierarchical organization of modularity in metabolic networks. Science 297(5586), 1551–1555 (2002)

Author Index